



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Your manuscript ID#21112502244011 has been received

2 messages

European Journal of Educational Research <editor@eu-jer.com>
Reply-To: European Journal of Educational Research <editor@eu-jer.com>
To: European Journal of Educational Research <sugeng.sutiarso@fkip.unila.ac.id>

Thu, Nov 25, 2021 at 9:45 AM

Dear Dr. Sugeng Sutiarso (sugeng.sutiarso@fkip.unila.ac.id),

Your manuscript entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" (ID#21112502244011) has been submitted successfully.

The link of your manuscript: https://eu-jer.com/aa/lib/elfinder/files/21112502244011/MS_EUJER_ID_21112502244011.docx

We will inform you about the developments of your paper in a month. Thank you for your interest to our journal.

Best regards.

Ahmet Savas, Ph.D.

Editor, European Journal of Educational Research
www.eu-jer.com
editor@eu-jer.com

European Journal of Educational Research <editor@eu-jer.com>
Reply-To: European Journal of Educational Research <editor@eu-jer.com>
To: European Journal of Educational Research <sugeng.sutiarso@fkip.unila.ac.id>

Thu, Nov 25, 2021 at 9:54 AM

Dear Dr. Sugeng Sutiarso (sugeng.sutiarso@fkip.unila.ac.id),

Your manuscript entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" (ID#21112502244011) has been submitted successfully.

The link of your manuscript: https://eu-jer.com/aa/lib/elfinder/files/21112502244011/MS_EUJER_ID_21112502244011_1.docx

[Quoted text hidden]



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Completed the preliminary review the manuscript EU-JER ID#21112502244011

1 message

European Journal of Educational Research <editor@eu-jer.com>
Reply-To: European Journal of Educational Research <editor@eu-jer.com>
To: European Journal of Educational Research <sugeng.sutiarso@fkip.unila.ac.id>
Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Tue, Dec 21, 2021 at 3:12 PM

Dear Dr. Sugeng Sutiarso,

Congratulations! Your paper has passed the test of plagiarism. We have completed the preliminary review for your manuscript entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" (Manuscript EU-JER ID#21112502244011). It is suitable for our journal's scope. We have sent your paper to the referees to evaluate.

We will inform you about the result, when we get the reports from referees.

PS: As you can see in our web site, we kindly remind that the authors were not allowed to withdraw submitted manuscripts after preliminary review because the withdrawal is a waste of valuable resources that editors and referees spent a great deal of time processing submitted manuscript, money, and works invested by the publisher.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Corrections request for the manuscript ID# 21112502244011

6 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: sugeng.sutiarso@fkip.unila.ac.id
Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Wed, Feb 2, 2022 at 10:22 PM

Dear Dr. Sugeng Sutiarso (sugeng.sutiarso@fkip.unila.ac.id),

Congratulations! After a thorough double-blind review, I am pleased to inform you that your manuscript entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" (ID#21112502244011) can be published on condition that corrections are made.

Please consider the reviewers' reports and emendations about your paper, please edit your manuscript and resend it as author names **blinded** paper by email attachment to us as soon as possible. In addition, we request to fill out the attached correction report what you have done as a word file. Please also highlight the edited parts in different (yellow and green) colors for each reviewer.

After we check your manuscript, we will send you the acceptance letter. The deadline for sending your finalized paper is **February 15, 2022** in order to publish in our next issue. If you need more time, please don't hesitate to contact me.

1- Please check the language of the whole paper as a proofreading lastly.

2- Please check all references for compatibility to APA 7 style (see <https://eu-jer.com/citation-guide>). Also please provide all issue, doi or nondatabase article link -if any (To find the DOI easily see: <http://doi.crossref.org/simpleTextQuery>).

3- Please provide English translation of the title of non English sources as at the below:

Eg.

Bussieres, E.-L., St-Germain, A., Dube, M., & Richard, M.-C. (2017). Efficacite et efficience des programmes de transition a la vie adulte: Une revue systematique [Effectiveness and efficiency of adult transition programs: A systematic review]. *Canadian Psychology/ Psychologie Canadienne*, 58(1), 354–365. <https://doi.org/10.1037/cap0000104>

Note for this example that "Canadian Psychology/ Psychologie Canadienne" is a bilingual journal that is published with a bilingual title; if the journal title were only in French it would not be necessary to translate it in the reference.

PS: If all of the corrections don't be completed, the paper can not be published. If you object to any correction, please explain this in your correction report.

Please **confirm** when you get this email. We are looking forward to getting your revised paper and correction report by email.

Best regards,

Ahmet Savas, Ph.D.

Editor, European Journal of Educational Research

editor@eu-jer.comwww.eu-jer.com

4 attachments

 **CORRECTION REPORT_EU-JER.docx**
19K

 **EU-JER_REVIEWER_FORM_R2612.docx**
288K

 **EU-JER_REVIEWER_FORM_R2613.docx**
287K

CORRECTION REPORT			
No	Reviewer Code	Reviews	Corrections made by the author
1			
2			
3			
4			
5			
6			
7			



European Journal of Educational Research

ISSN: 2165-8714

<http://www.eu-jer.com/>

Review Form

Manuscript ID: MS_EUJER_ID_21112502244011 **Date:** February 2, 2021

Manuscript Title: **Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School**

ABOUT MANUSCRIPT (Mark with "X" one of the options)	Accept	Weak	Refuse	Not Available
Language is clear and correct		X		
Literature is well written			X	
References are cited as directed by APA	X			
The research topic is significant to the field		X		
The article is complete, well organized and clearly written		X		
Research design and method is appropriate		X		
Analyses are appropriate to the research question		X		
Results are clearly presented		X		
A reasonable discussion of the results is presented		X		
Conclusions are clearly stated		X		
Recommendations are clearly stated		X		

GENERAL REMARKS AND RECOMMENDATIONS TO THE AUTHOR

This study aims to develop an assessment test using polytomous item response theory. However, the introduction is overly long and confusing. There is much information about the item response theories. However, most of this information should be moved to the method. Instead, in the introduction, the authors should focus on the strengths and weaknesses of the existing instruments that developed to assess students' learning in mathematics in the context of vocational/high school (?) students. The current version of the introduction opens many questions. These are: What do we know about the existing instruments? What do we need to know about the existing instruments? Why is this study important for students and teachers, and scholars? These need to be answered in this section.

Method

More information regarding participants should be given. For example, we need to have information about their grade, grade level, and gender.

How many experts were enrolled to assess the validity? Please give details. In addition, what were the experts' feedbacks on the draft version of the instrument?

It is not clear how the polytomous item response theory was implemented in the method and the results. More details are needed.

Results

Figure 3 is not addressed in the text.

Discussion

The authors did write the following sentence: "One of the reasons that make this instrument suitable for use in the process of preparing the instrument under expert supervision." However, I do not agree with the comment of the authors as the expert view has been received for all developed instruments.



European Journal of Educational Research

ISSN: 2165-8714

<http://www.eu-jer.com/>

The discussion is overly long and hard to read and follow the text. The authors did still repeat the results in this section. This repetition of the results makes the text hard to understand and misses the focus of the study. Therefore, I suggest that the authors should make a brief overview of the results. Later, they should discuss similarities between the existing instruments in the literature and the developed instrument in this study. After this, they should discuss the differences between previous studies and this study. While discussing the similarities and differences, the author should also discuss possible reasons for the results.

The results of analyzing students' answers must be moved to the results. In the discussion, we need the results themselves.

Conclusion

What is the new knowledge from this research for researchers and the literature? Please explain this detail.

Recommendations

Please make more specific suggestions for research in the future.

Language

The use of the language is very problematic. The text needs proofreading by a native speaker.

THE DECISION (Mark with "X" one of the options)

Accepted: Correction not required	
Accepted: Minor correction required	
Conditionally Accepted: Major Correction Required (Need second review after corrections)	X
Refused	

Reviewer Code: R2612 (The name of referee is hidden because of blind review)




European Journal of Educational Research

ISSN: 2165-8714

<http://www.eu-jer.com/>

Review Form

Manuscript ID:	EU-JER_ID#21112502244011	Date: 13/01/2022			
ManuscriptTitle:	Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School				
ABOUT MANUSCRIPT (Mark with "X" one of the options)		Accept	Weak	Refuse	Not Available
Language is clear and correct			X		
Literature is well written			X		
References are cited as directed by APA		X			
The research topic is significant to the field		X			
The article is complete, well organized and clearly written		X			
Research design and method is appropriate		X			
Analyses are appropriate to the research question		X			
Results are clearly presented		X			
A reasonable discussion of the results is presented			X		
Conclusions are clearly stated		X			
Recommendations are clearly stated			X		
GENERAL REMARKS AND RECOMMENDATIONS TO THE AUTHOR					
<p>Discussion section seems like more detailed findings section. Use previous studies to compare or contrast your results in Discussion section.</p> <p>Add recommendations for future research to Recommendations section.</p>					
THE DECISION (Mark with "X" one of the options)					
Accepted: Correction not required					
Accepted: Minor correction required					X
Conditionally Accepted: Major Correction Required (Need second review after corrections)					
Refused					
Reviewer Code: R2613 (The name of referee is hidden because of blind review)					

 MS_EUJER_ID_21112502244011_R2613.docx
540K

SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Thu, Feb 3, 2022 at 3:27 PM

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research

I am very happy to hear this news. I and my colleagues will revise our manuscript soon within the deadline.

Best regards,
Sugeng Sutiarso

On 2/2/22, Editor - European Journal of Educational Research
<editor@eu-jer.com> wrote:

> Dear Dr. Sugeng Sutiarso (sugeng.sutiarso@fkip.unila.ac.id),
>
> Congratulations! After a thorough double-blind review, I am pleased to
> inform you that your manuscript titled "Development of Mathematics
> Assessment Instruments for Learning with Polytomous Response in
> Vocational School" (ID#21112502244011) can be published on condition
> that corrections are made.
>
> Please consider the reviewers' reports and emendations about your paper,
> please edit your manuscript and resend it as author names *blinded*
> paper by email attachment to us as soon as possible. In addition, we
> request to fill out the attached correction report what you have done as
> a word file. Please also highlight the edited parts in different (yellow
> and green) colors for each reviewer.
>
> After we check your manuscript, we will send you the acceptance letter.
> The deadline for sending your finalized paper is *February 15, 2022* in
> order to publish in our next issue. If you need more time, please don't
> hesitate to contact me.
>
> 1- Please check the language of the whole paper as a proofreading lastly.
>
> 2- Please check all references for compatibility to APA 7 style (see
> <https://eu-jer.com/citation-guide>). Also please provide all issue, doi
> or nondatabase article link -if any (To find the DOI easily see:
> <http://doi.crossref.org/simpleTextQuery>).
>
> 3- Please provide English translation of the title of non English
> sources as at the below:
>
> Eg.
>
> Bussieres, E.-L., St-Germain, A., Dube, M., & Richard, M.-C. (2017).
> Efficacite et efficience des programmes de transition a la vie adulte:
> Une revue systematique [Effectiveness and efficiency of adult transition
> programs: A systematic review]. /Canadian Psychology/ Psychologie
> Canadienne, 58/(1), 354–365. <https://doi.org/10.1037/cap0000104>
>
> Note for this example that "Canadian Psychology/ Psychologie Canadienne"
> is a bilingual journal that is published with a bilingual title; if the
> journal title were only in French it would not be necessary to translate
> it in the reference.
>
> PS: If all of the corrections don't be completed, the paper can not be
> published. If you object to any correction, please explain this in your
> correction report.
>
> Please *confirm* when you get this email. We are looking forward to
[Quoted text hidden]

Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School

Abstract: This study aims to produce an assessment for learning instruments with a polytomous response. This research is development research that refers to the general education development model of Plomp. The sample of this study was students of Vocational School in East Lampung Regency, Indonesia with a total of 413 students. The research data were collected using multiple-choice tests with open reasons. The instrument consists of two package questions with each package consisting of 40 items. The item validity index was analyzed using the Gregory index formula and the gain parameters were analyzed using the Partial Credit Model 1-Parameter Logistic approach (1-PL) using the Winsteps program. The results of the study indicate that the instrument has been valid. This is based on the calculation that the Gregory index value is 1. In addition, the average reliability coefficient of 0.89 is obtained, so it is stated reliable to measure the ability of students. The measurement results in terms of the ability of students to answer each item are stated in the range of logit -2 to +2; if it is viewed from the percentage of student mastery in answering each item of the questions, the instrument is stated in the medium category.

Keywords: assessment for learning instrument, open-reasoned multiple-choice, polytomous response

Introduction

Assessment is important to determine the extent to which the teaching and learning process is by the stated goals William (2011). One type of assessment that is important to carry out is assessment for learning, or often referred to as formative assessment. With the assessment of an educator, it is easy to assess the extent to which students' understanding and mastery of concepts and obtain information on the extent to which students can apply, and analyze the material taught by educators. Educators can evaluate learning outcomes by giving a test in the process of teaching and learning activities. A test is a tool or procedure which is used to know or measure something in an atmosphere, using rules that have been determined (Arikunto, 2012). The purpose of conducting the test is to find out the learning achievement or

competency of students and can provide information about the cognitive ability or students' skills.

Multiple choice test is one form of selected response test that is widely used for various purposes (Haladyna, 2004). This is inseparable from the superiority of effective multiple-choice tests to measure various types of knowledge and complex learning outcomes, it is very appropriate for exams with many participants and the results must be announced immediately. But apparently, there are some weaknesses, such as students do not have the freedom to write, organize, and express their ideas that are expressed in their own words or sentences; it cannot be used to measure problem-solving skills; it is very sensitive to guessing; preparation of good tests requires a relatively long time compared to other forms of tests; and it is very difficult to determine alternative answers (distractors) that are truly homogeneous, logical, and functioning.

Each form of multiple-choice questions and essays each have advantages and disadvantages, namely, the form of multiple-choice questions about goals and the complexity of time cannot describe the ability of students basically (Roediger III & Marsh, 2005). The form of essay questions can describe the true abilities of students, but the examiners seem subjective (Ellery, 2008). Based on the advantages and disadvantages of the 2 forms of the question, in the study will be developed assessment instruments for learning in the form of reasoned multiple-choice to still accommodate the two forms of the question.

The beginning of the use of reasoned multiple-choice tests began in the 80s which is aimed to identify student misconceptions. However, so far there has been no research that develops multiple-choice questions with reasons in the field of mathematics, and even if there is research outside of mathematics. The research is the development of multiple-choice questions with reasons in the field of chemistry (Krishnan & Howe, 1994), the field of

English (Williams, 2006), and the field of economics (Buckles & Siegfried, 2006). Research in the field of mathematics is only limited to developing multiple-choice questions without reason (Torres, Lopes, Azevedo, & Babo, 2009).

The reasoned multiple choice test consists of two types: open-reasoned multiple-choice test and closed-reasoned multiple-choice test (Hirsch & O'Donnell, 2001). An open-reasoned multiple-choice test is a test accompanied by reasons so students must write down the reasons for the answers chosen. The advantage of an open-ended multiple-choice test is students can freely express the reasons for their chosen answers. The disadvantage is it takes time for understanding students' broad answers. While the closed-multiple choice test is a multiple-choice test accompanied by a choice of reasons. This test is also called a two-tier multiple-choice test. The first level is a multiple choice question with an answer choice while the second level is a multiple choice with a choice of reasons for the answer at the first level. Chandrasegaran, Treagust, and Mocerino (2007) said that the student's reasons in closed-choice multiple-choice forms have been provided so that students only choose answers from the available options. The correct answer is if students correctly choose the option at the first and second levels. The weakness of this instrument is students are not free to express the reason for choosing an answer. The advantage of this instrument is it simplifies the assessment process. In addition, students have the opportunity to guess answers smaller than one level multiple choice.

Student responses consist of two parts, namely dichotomous and polytomous (Whitehead, Huang, Blomquist, & Ready, 1998). In the dichotomous item the response to answers there is only two possibilities, like true or false, yes or no, with a value of zero or one. The item whose response is more than one possibility is called the polytomous item. Polytomous item response models can be categorized as nominal and ordinal item response models, depending

on the characteristic assumptions about the data. The nominal item response model can be applied to items that have an ordered herd alternative and the various levels of capability measured. The ordinal response model occurs in items that can be scaled into the number of certain categories arranged in the answer.

According to Hambleton, Swaminathan, and Rogers (1991) there are three assumptions in the item response theory, they are;

1. Unidimensional means that there is only one ability measured by a set of items in the test,
2. Local Independence means that when the ability to influence performance remains constant, the response of the examinee to each pair of items is statistically independent of each other,
3. The invariance of capability parameters is the basis of IRT and the main difference from classical test theory.

There are important things that need to be considered in the item response theory, it is the selection of the right model. The right model will reveal the true state of the test data as a measurement result. There are 3 models of relationships between abilities and grain parameters, they are (Hambleton & Jones, 1993);

1. Model parameter-1(Rasch model), determined by the item difficulty index (bi).
2. Model parameter-2, determined by the item difficulty level index (bi) and grain difference power index (ai).
3. Model parameter-3, determined by the item difficulty level index (bi), grain difference power index (ai), and false guesses (ci).

Based on various scoring models and parameter models, the formula that will be used for analyzing this research is the Partial Credit Model/PCM (Masters, 2011). The PCM model is suitable for analyzing test items that require several steps of completion, this includes

mathematical questions that require the identification phase of the problem to the final solution. PCM is an extension of the Rasch model, assuming that each item has the same difference power. If i is a polytomy item with a scoring category of 0, 1, 2, ..., m_i , then the probability of individual n scores x in item i is described in the category response function (CRF) and realized in equation 2 (Ostini & Nering, 2006).

Five characteristics need to be considered in the modern response theory, they are;

1. The level of difficulty (trait)

According to DeMars (2010), the level of difficulty of the item identifies an ability where around 50% of the examinees (or fewer, depending on the model) are expected to answer the item correctly. In theory, the b_i value is located between $- \sim$ and $+ \sim$. Hambleton, Swaminathan, and Rogers (1985) stated that an item is said to be good if this value ranges between -2 and $+2$. Supporting this statement, Retnawati (2014) revealed that if the b_i value approaches -2 , then the index of grain difficulties is very low. Whereas, if the b_i value is close to $+2$, the item difficulty index is very high for a group of test-takers. The criteria for items that are good are not too easy and not too difficult. With varying levels of difficulty, it can measure the ability of the test takers as a whole. So that the greater the index of the difficulty of the item, the more difficult the item questions to do, on the contrary, the smaller the index of the difficulty of the item, the easier the item will be.

2. Difference (slope)

The index usually shows how steep the possibility of correct response changes such as ability or changes in trait. According to DeMars (2010) that higher differentiating power means that the item can distinguish (discriminate) between examinees with various levels of constructs. On the characteristic curve, a_i is (slope) of the curve at the point b_i for a certain scale of ability, because it is a slope, the greater the slope, the greater the distinguishing power of the item. In theory, Retnawati (2014) stated that the value of a_i

value lies between $- \sim$ and $+ \sim$. The more items that reach different power criteria ($a-i$), the test items are better distinguishing test takers' abilities.

3. Match Points with Logistics Models (statistical goodness of fit)

According to Retnawati (2014), the suitability of the model can be known by comparing chi-square (χ^2) tables with certain degrees of freedom. The item is said to fit a model if the calculated chi-square value does not exceed the chi-square value of the table. Compatibility can also be known from the probability value (significance, sig). If the value is $\text{sig} < \alpha$, then the item is said not to match the model. The suitability of the model can also be seen by looking at the proportion of items that match the logistical model. The proportion of the most suitable items between the 1PL, 2PL, and 3PL models is expressed as a suitable model for the test item test. Another way to do this is to plot the characteristic curve. The plot can be illustrated with the help of the Winsteps program, with a plot to find out how precise the data distribution is compared to the model.

4. Value of Information Function

The information item function is a model for explaining the size of an item on the test device, selection of test items, and comparison of several test devices (Retnawati, 2014). The information function states the strength or contribution of test items in revealing the latent abilities measured by the test. Thus, through the item information function, it can be known which items match with the model so that it can assist in the selection of test items. The test information function is needed to interpret the test results. A good test tool will have an information function value that is greater than the measurement error. If the measurement error is higher than the information, it can be estimated that the test plan is not suitable for the ability of the participant given the test.

5. Standard Error of Measurement (SEM)

Standard error $\hat{\theta}$, $SE(\hat{\theta})$ is the asymptotic standard deviation of the normal distribution of the maximum likelihood estimation for the ability given to the actual value of ability θ . In item response theory, standard assessment errors of SEM (Standard Error of Measurement) are closely related to information functions. According to Hambleton et al. (1991) that the information function with SEM has an inversely proportional relationship, the greater the information function, the smaller the SEM or vice versa.

This study aims to produce an assessment for learning instruments with a polytomous response. The research questions posed are (1) how is the process of developing an assessment for learning an instrument with a polytomous response in mathematics subjects at the Vocational High School level?, and (2) does the developed instrument have the parameters of an assessment for learning an instrument with a quality polytomous response?.

Methodology

Research Design

This research is development research that refers to the general education development model of Plomp (2013). The development model proposed by Plomp consists of five stages, they are preliminary investigation, design, realization/construction, test, evaluation and revision, and implementation.

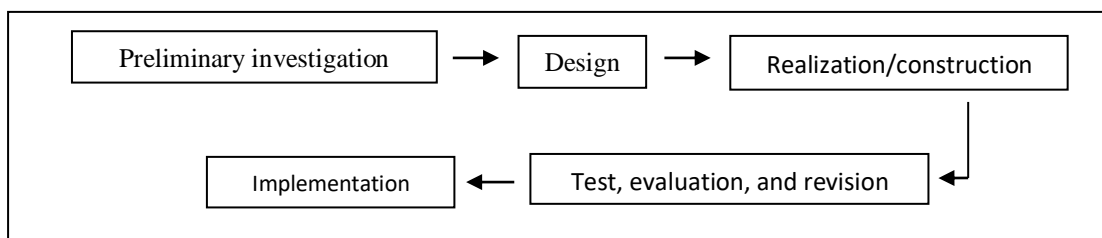


Figure 1. The Development Model of Research

Sample and Data Collection

The sample of this study was students of Vocational School (VS) in East Lampung Regency, Indonesia. In this case, it will be represented by 3 schools, namely: VS of Mitra Bhakti, Praja Utama, and Ma'arif with a total of 413 students.

Table 1. Sample of Research

School	Class	Responden
VS of Mitra Bhakti	X AK	13
	X PM1	23
	X PM2	29
VS of Praja Utama	X TKR1	40
	X TKR2	39
	X AK1	43
	X AK2	38
	X PM1	41
	X PM2	40
	X AP	40
VS of Ma'arif	X TKJ	31
	X AK	36
Total		413

The research data were collected using multiple-choice tests with open reasons. The instrument consists of two package questions with each package consisting of 40 items.

Analyzing of Data

Data analysis techniques used in this study are content validity, construct validity, reliability, level of difficulty of items, and analysis of the characteristics of test items in the form of general assumption tests, assumptions of local independence assumptions, item compatibility, and information and measurement errors. The technique for content validity is asking the expert, in this case as a validator to check the accuracy and give an assessment of the suitability of the item with the indicators, the question writing editor, and the suitability of the choice choices (deception) in multiple choices. The assessment can be seen in Table 2.

Table 2. Criteria for Grading Items by Experts

Score	Description
1	It is not relevant
2	It is relevant enough
3	It is relevant
4	It is very relevant

After being assessed by an expert, the researcher calculates the results of the assessment using an index of validity, including the index proposed by Gregory (Retnawati, 2016a). With a range of V numbers that may be obtained from 0 to 1. The higher the V number (close to 1 or equal to 1), the higher the value of the validity of an item, and the lower the V number (close to 0 or equal to 0) then the value of the validity of an item is also getting lower. How to calculate the Gregory Index is as follows:

Table 3. Calculating Gregory Index

		Rater 1	
		Low	High
Rater 2	Low	A	B
	High	C	D

With content validity coefficient: $V = \frac{D}{A+B+C+D} = 1$

Description:

V: The validity

A: Validators 1 and 2 who rate low

B: Validator 1 who rate high, but validator 2 who rate low

C: Validator 1 who rate low, but validator 2 who rate high

D: Validators 1 and 2 who rate high

In addition, to prove construct validity, exploratory factor analysis was used. Exploratory factor analysis can be seen from the value of KMO (Kaiser Meyer Olkin). KMO value is obtained through IBM SPSS 20. 20. KMO value **More** than 0.5 indicates the variables and samples used to allow further analysis (Retnawati, 2016b). Meanwhile, instrument reliability was estimated using internal consistency techniques with the **Cronbach-alpha** formula which was assisted by IBM SPSS 20. Cronbach's Alpha value 0.6 and less than 1 indicated that the instrument met reliable criteria, whereas if Cronbach's Alpha value was less than 0.5 shows

the instrument is not reliable. Furthermore, the difficulty level of the item (D) in the form of reasoned multiple choices can be calculated using a formula (Nitko, 1996) is:

$$D = \frac{S}{T}$$

Description:

D: The difficulty level of the item

S: The number of students who answered correctly

T: Number of students taking the test

After doing the calculation, the items can be categorized into items that are very easy, moderate, difficult, and very difficult to refer to Baker (2001). Here is a table of difficulty levels:

Table 4. Categories of Difficulty Levels

Very Easy	Easy	Moderat	Difficult	Very Difficult
- 2.0	- 0.5	0	+0.5	2.0

Furthermore, the data analysis begins by describing the feasibility of the characteristics of the vocational multiple-choice objective test using theory response items with the help of the Winstep Program. Winstep program is used because it has several advantages (Subali & Suyata, 2011), they are (1) can analyze data in the form of dichotomous and polytomous, and (2) the availability of the results of modern theory analysis is based on the maximum likelihood model using a one-parameter logistic model.

Analysis using IRT can be done by testing the dimensional assumptions through analysis of fit or explanatory factor analysis. Test items are carried out in dimensions if the item measures one ability. If the unidimensional assumptions have been fulfilled, then automatically the local independence assumptions have also been fulfilled. Indications that the items are unidimensional are data fit with the model. To find out whether the model used is by the item, it can be used mean-Square (IMS) and Mean-Square Outfit (CSO) statistics.

IMS and CSO statistics are the degree of conformity between observation data and predictive values by the model. Test items are said to be fit models if they have IMS and CSO values ranging from 0.5 to 1.5 (Hanafi, Ab Rahman, Mukhtar, Ahmad, & Warman, 2014).

Findings / Results

Content validity

The study of the instruments carried out by 2 experts in mathematics learning produced valid proof of the instrument contents. The results of the study show that the instruments made by researchers are in good category and can be used but there are still some things that need to be corrected. The researcher repairs/revise as much as possible, the suggestion is written by the validator on the instrument sheet. The results of the assessment are analyzed using the Gregory index formula to find out about each item's problems.

Table 5. Results of Item Assessment by the Expert

	Rater 1	
Rater	0	0
2	0	40

Based on the results of the assessment presented in Table 5, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Because each item about the instrument meets the valid criteria, the instrument is ready to be tested. The researcher did not test all the instruments that had been made. The question instrument which was tested was 40 items. In addition, the instrument was piloted in schools that have used curriculum 13.

Construct validity

After completing the trial research on the problem in the field, the researchers then conducted a scoring activity. This is done to prove the construct validity using exploratory factor analysis. The results of the exploratory factor analysis can be seen in Table 6.

Table 6. Results of Exploratory Factor Analysis

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	0.000

Based on Table 6, it can be seen that the KMO value can be explained more than 0.5. Therefore, it can be concluded that the variables and samples used allow further analysis.

Instrument reliability

The estimation of the instruments developed that is showed satisfactory results. Based on the results of the instrument reliability estimation in Table 7, information is obtained that the Cronbach's alpha coefficient in the test package tested is 0.89, which means the package is reliable. So that the question package can be used for further research.

Table 7. Instrument Reliability

Cronbach's Alpha	N of Items
0.892	40

Test of unidimensional assumptions

The unidimensional assumption test is the first assumption that must be fulfilled in using item response theory analysis. The unidimensional test was seen based on the cumulative percentage of eigenvalues and scree plots from the analysis using the SPSS program. The results of the analysis can be seen in Figure 2. Based on the scree plot, it is known that the value of the eigenvalue is directly sloping on the second factor, this shows that there is only 1 dominant factor. These results prove that this test device meets the dimensional assumptions.

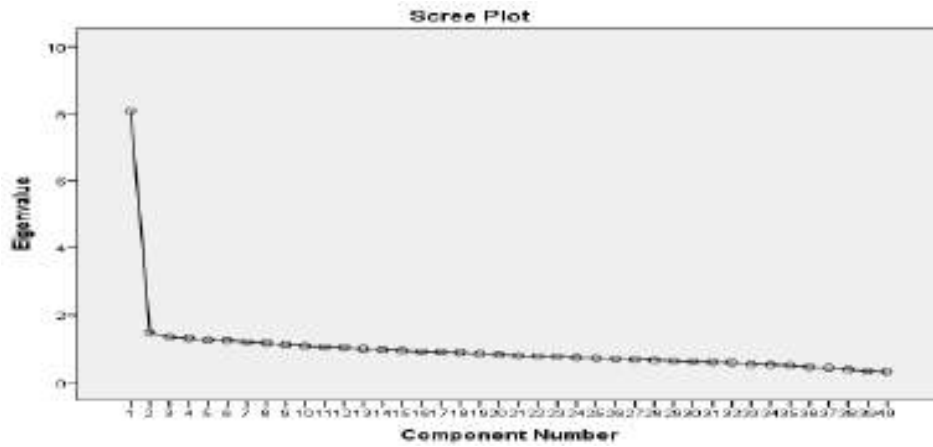


Figure 2. Scree Unidimensional Plot

Test local independence assumptions

Testing the assumption of local independence is done through analysis with the Winsteps program. After obtaining personal measures from the Winsteps program, the next analysis is done with the Microsoft Excel program and the results can be seen in Table 8.

Table 8. Covariance Matrix

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Based on Table 8, it can be seen that the covariance value between the ability interval groups located in the diagonals is small and close to zero, this can be concluded that there is no correlation, so that the assumption of local independence assumption is fulfilled.

Match item (Fit item)

The item compatibility test assuming the Rasch Model approach is done by looking at the fit or not of the items on the model. This test is analyzed using the Winsteps program. Based on the results of the analysis, of the 40 items that have been made, all match the model or fit. The results of the analysis are shown in Table 9.

Table 9. Match Item to Model

Item	Outfit MNSQ	Pt-Measure Corr	Description	Item	Outfit MNSQ	Pt-Measure Corr	Description
1	1.16	0.74	Model fit	21	1.02	0.44	Model fit
8	1.16	0.51	Model fit	14	1.01	0.42	Model fit
6	1.16	0.44	Model fit	25	1.00	0.41	Model fit
32	1.12	0.35	Model fit	2	0.98	0.75	Model fit
18	1.12	0.35	Model fit	11	0.98	0.38	Model fit
15	1.10	0.49	Model fit	20	0.97	0.47	Model fit
16	1.08	0.44	Model fit	26	0.97	0.31	Model fit
22	1.08	0.56	Model fit	9	0.97	0.43	Model fit
31	1.07	0.23	Model fit	29	0.97	0.33	Model fit
24	1.07	0.37	Model fit	37	0.96	0.40	Model fit
13	1.07	0.44	Model fit	17	0.96	0.48	Model fit
34	1.07	0.33	Model fit	3	0.94	0.67	Model fit
30	1.05	0.30	Model fit	10	0.93	0.40	Model fit
23	1.05	0.53	Model fit	12	0.90	0.39	Model fit
19	1.05	0.26	Model fit	38	0.89	0.43	Model fit
36	1.05	0.34	Model fit	28	0.88	0.40	Model fit
35	1.04	0.40	Model fit	5	0.84	0.41	Model fit
33	1.04	0.41	Model fit	4	0.82	0.46	Model fit
7	1.03	0.45	Model fit	40	0.74	0.54	Model fit
27	1.02	0.32	Model fit	39	0.66	0.59	Model fit

Difficulty level (Item difficulty)

Analysis of the level of difficulty of the items is done with the Winsteps program and the results obtained can be presented in table 10. It can be seen that the range was obtained from - 0.70 to 0.84.

Table 10. Level of Difficulty of Item

Item	Total Score	Degree of difficulty	Category	Item	Total Score	Degree of difficulty	Category
40	801	0.84	Middle	28	975	0.02	Middle
39	925	0.72	Middle	32	974	0.02	Middle
38	921	0.27	Middle	7	977	0.01	Middle

Item	Total Score	Degree of difficulty	Category	Item	Total Score	Degree of difficulty	Category
27	936	0.2	Middle	29	981	-0.01	Middle
26	943	0.16	Middle	33	981	-0.01	Middle
30	949	0.14	Middle	11	986	-0.03	Middle
31	955	0.11	Middle	12	988	-0.04	Middle
35	955	0.11	Middle	16	994	-0.07	Middle
37	955	0.11	Middle	20	997	-0.08	Middle
13	957	0.1	Middle	6	998	-0.09	Middle
34	959	0.09	Middle	15	1000	-0.09	Middle
36	961	0.08	Middle	14	1003	-0.11	Middle
10	964	0.07	Middle	21	1010	-0.14	Middle
17	964	0.07	Middle	5	1021	-0.19	Middle
8	966	0.06	Middle	23	1037	-0.26	Middle
9	965	0.06	Middle	22	1050	-0.32	Middle
25	968	0.05	Middle	4	1064	-0.38	Middle
18	972	0.03	Middle	3	1066	-0.39	Middle
19	974	0.02	Middle	2	1079	-0.45	Middle
24	974	0.02	Middle	1	1134	-0.70	Middle

Information function and measurement errors

The information function is inversely proportional to SEM (Standard Error Measurement).

The following is a picture of the curve of the relationship between information functions and measurement errors.

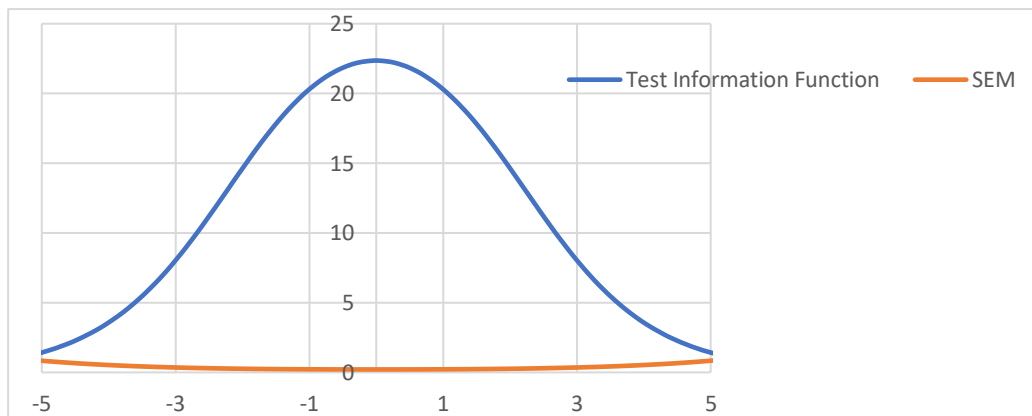


Figure 3. Graph of Information Functions and Measurement Errors

With the item information function, it is known which items match the model, so that it helps in the selection of test items. The graph indicates that the greater the value of the information function, the smaller the error rate (SEM). Conclusion The characteristics of the test device are suitable for students with moderate abilities.

Discussion

The results of this study indicate that the learning assessment instrument can be used in vocational school students. These results are very encouraging because since the beginning of the assessment of multiple-choice instruments that are reasoned in the field of mathematics have not existed (no research). One of the reasons that make this instrument suitable for use in the process of preparing the instrument under expert supervision. These expert suggestions have provided an assessment of the instrument of appropriateness and direction that is correct and by the circumstances of the research subject. Much has been done on the assessment of the instrument, both in terms of material and the rules of the language used. (Çanakkale & Çanakkale, 2013) said that To ensure the validity of the content, it is necessary to have a review from an expert. In addition, many students who were involved in the trial (413 people) exceeding the minimum limit of 200 people have given confidence that this instrument is suitable for use (Malhotra, Nunan, & Birks, 2017).

Procedure for Development of Assessment for Learning Instruments

The development of an assessment for learning an instrument with a polytomous response has gone through a series of development stages with the development model from Plomp to produce a product. Based on the results of the instrument content validation analysis conducted by the validator. It is known that the assessment for learning an instrument with a polytomous response that has been developed is valid with a Gregory index of 1. This means that the items on the test instrument can be used to measure the level of student mastery (Suhaini, Ahmad, & Bohari, 2021)

Based on the results of the reliability analysis of the assessment for learning instrument using SBM SPSS, it is known that the instrument is classified as reliable with a value of 0.89 (high).

This means that the assessment for learning instrument that has been developed can be trusted and gives the same results if the test is carried out on different subjects, places, and conditions. Based on the results of the analysis of the assumptions or parameters of the item response theory. The analysis was performed using IBM SPSS software, Ms. Excell, and Winsteps obtained results indicating that the assessment for the learning instrument had met the requirements.

The first assumption test that is fulfilled is unidimensional, which is seen in the cumulative percentage of the first eigenvalue more than 20% and is indicated by a scree plot that slopes directly on the second factor. The second condition in item response theory that is fulfilled is local independence. Based on the results of the analysis with the Winsteps program and continued with Ms. Excell, it can be seen that the covariance value between groups is close to zero. This means that there is no correlation, so it can be said that the local independence assumption test is fulfilled.

Based on the results of the analysis of the suitability of the items using the Winsteps program, the Outfit MNSQ was 0.5 to 1.5 and the Pt-Measure Corr was positive. This means the item match meets the fit criteria with the Rasch model (1 Logistics Parameter). Based on the parameter analysis of the item difficulty level using the Winsteps program, it was found that the instrument was at a moderate level of difficulty. The level of difficulty can be seen from the parameters that are still in the range of -0.70 to 0.84. Thus all items meet the criteria for a good level of difficulty. Based on the results of the analysis of the information function and measurement error, the test instrument can be given to students with moderate abilities. These results can be seen from the graph of the results of data processing by Ms. Excel. In these results, it can be seen that the information value is shown at 22.36 and the measurement error is 0.211.

Results of Analysis of Assessment for Learning Instruments

Assessment for learning is an assessment that occurs during the learning process that involves interactive teachers with students or students with other students to foster student motivation to enhance activities in learning. Then is made a joint decision between teachers and students to create an atmosphere of further learning and is carried out continuously to achieve student learning development in the dimensions of knowledge, attitudes, and skills. Assessment can be used as a standard reference for student success in achieving learning outcomes (Syaifuddin, 2020). The assessment can also be used as a consideration to make a decision in the learning process, to achieve better learning.

Success in learning through assessment is intended for both teachers and students. Teachers are required to have adequate insight and abilities about learning, for example, planning, setting learning goals, and making the right decisions based on the information obtained in the assessment, so that students are motivated to improve and improve their learning. Then, the assessment for learning also provides insight about learning to students, that all students have the opportunity to achieve success in learning.

Based on the results of field trials, in addition to knowing the quality of the test instruments developed, it can also be seen the ability of students to work on the questions given. Instrument analysis is based on two patterns of student answers that have the same tendency. A total of 6 students' answers were selected for further analysis by taking into account the students' abilities, namely two students of each different ability (high, medium, and low). This analysis is based on Bloom's Taxonomy of learning (Bloom, 1956), and the results of the analysis are as follows.

Question 1:	Pattern 1:	Pattern 2:
-------------	------------	------------

<p>Diketahui barisan bilangan 3, 8, 13, 18, Rumus suku ke-n barisan itu adalah ...</p> <p>a. $U_n = 5n - 3$ b. $U_n = 5n - 2$ c. $U_n = 2n + 1$ d. $U_n = 4n - 1$ e. $U_n = 3n + 2$</p>	<p>1. diketahui : $a = 3$ $b = 5$ ditanya $U_n = ?$ jawab : $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	<p>$u_1 = a = 3$ $b = u_2 - u_1 = 8 - 3 = 5$ $u_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 4. Students' Answers to Question 1

In question 1, the cognitive domain to be achieved is C2 (Understanding). Based on students' answers, that the two dominant patterns of student answers are (1) students have understood the general form of arithmetic sequences and know the first and different terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms of arithmetic sequences.

Question2:	Pattern 1	Pattern 2
<p>Diketahui barisan aritmetika 2, 5, 8, 11, 68. Banyaknya suku barisan tersebut adalah ...</p> <p>a. 12 b. 13 c. 22 d. 23 e. 24</p>	<p>2, 5, 8, 11, ... 68 0 2, 5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 59, 62 65, 68 Jadi banyak suku tersebut adalah 23</p>	<p>$68 = a + (n-1)b$ $68 = 2 + (n-1) \cdot 3$ $68 - 2 = (n-1) \cdot 3$ $66 = (n-1) \cdot 3$ $21 \cdot 3 = 1 \cdot 3 + n \cdot 3$ $n = 22 + 1$ $n = 23$</p>

Figure 5. Students' Answers to Question 2

In question 2, the cognitive domain to be achieved is C2 (understanding). Based on students' answers, that the two dominant patterns of student answers are (1) being able to understand and be able to determine the number of terms in a sequence by using the general formula for an arithmetic sequence or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term). to the last term), and (2) students who have been able to use general formulas for arithmetic sequences but have not been able to determine the number of arithmetic sequences due to errors in performing algebraic operations on general arithmetic sequences.

Question 3:	Pattern 1	Pattern 2
-------------	-----------	-----------

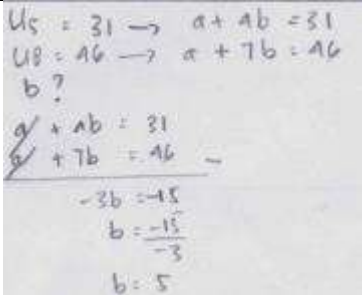
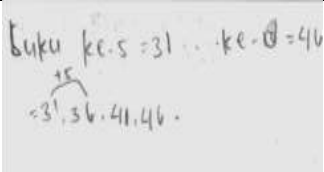
<p>Pada suatu barisan aritmetika diketahui suku ke-5 adalah 31 dan suku ke-8 adalah 46. Nilai beda barisan tersebut adalah ...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p>	 <p> $U_5 = 31 \rightarrow a + 4b = 31$ $U_8 = 46 \rightarrow a + 7b = 46$ $b ?$ $\frac{a + 4b = 31}{a + 7b = 46} -$ $-3b = -15$ $b = \frac{-15}{-3}$ $b = 5$ </p>	 <p> buku ke-5 = 31 .. ke-8 = 46 \uparrow +5 = 31, 36, 41, 46 </p>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 6. Students' Answers to Question 3

In question 3, the cognitive domain to be achieved is C3 (applying). Based on students' answers, that the two dominant patterns of student answers are (1) students have understood and can determine the difference or difference in an arithmetic sequence of two non-adjacent terms using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even if they do not use a general formula or by writing the terms from known terms and inserting several terms.

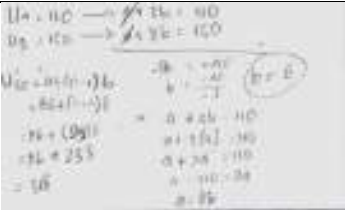
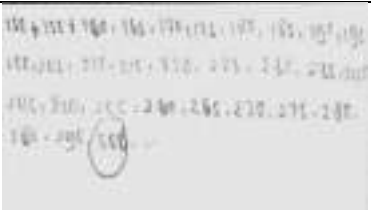
Question 4:	Pattern 1	Pattern 2
<p>berturut-turut adalah 110 dan 150. Suku ke-30 barisan aritmetika tersebut adalah ...</p> <p>a. 308 b. 318 c. 326 d. 344 e. 354</p>	 <p> $U_1 = 110 \rightarrow a = 110$ $U_2 = 150 \rightarrow a + b = 150$ $110 + b = 150$ $b = 40$ $U_{30} = a + (n-1)b$ $= 110 + (30-1)40$ $= 110 + 29 \cdot 40$ $= 110 + 1160$ $= 1270$ </p>	 <p> $110 + 110 + 40 = 150$ $150 + 40 = 190$ $190 + 40 = 230$ $230 + 40 = 270$ $270 + 40 = 310$ $310 + 40 = 350$ $350 + 40 = 390$ $390 + 40 = 430$ $430 + 40 = 470$ $470 + 40 = 510$ $510 + 40 = 550$ $550 + 40 = 590$ $590 + 40 = 630$ $630 + 40 = 670$ $670 + 40 = 710$ $710 + 40 = 750$ $750 + 40 = 790$ $790 + 40 = 830$ $830 + 40 = 870$ $870 + 40 = 910$ $910 + 40 = 950$ $950 + 40 = 990$ $990 + 40 = 1030$ $1030 + 40 = 1070$ $1070 + 40 = 1110$ $1110 + 40 = 1150$ $1150 + 40 = 1190$ $1190 + 40 = 1230$ $1230 + 40 = 1270$ $1270 + 40 = 1310$ $1310 + 40 = 1350$ $1350 + 40 = 1390$ $1390 + 40 = 1430$ $1430 + 40 = 1470$ $1470 + 40 = 1510$ $1510 + 40 = 1550$ $1550 + 40 = 1590$ $1590 + 40 = 1630$ $1630 + 40 = 1670$ $1670 + 40 = 1710$ $1710 + 40 = 1750$ $1750 + 40 = 1790$ $1790 + 40 = 1830$ $1830 + 40 = 1870$ $1870 + 40 = 1910$ $1910 + 40 = 1950$ $1950 + 40 = 1990$ $1990 + 40 = 2030$ $2030 + 40 = 2070$ $2070 + 40 = 2110$ $2110 + 40 = 2150$ $2150 + 40 = 2190$ $2190 + 40 = 2230$ $2230 + 40 = 2270$ $2270 + 40 = 2310$ $2310 + 40 = 2350$ $2350 + 40 = 2390$ $2390 + 40 = 2430$ $2430 + 40 = 2470$ $2470 + 40 = 2510$ $2510 + 40 = 2550$ $2550 + 40 = 2590$ $2590 + 40 = 2630$ $2630 + 40 = 2670$ $2670 + 40 = 2710$ $2710 + 40 = 2750$ $2750 + 40 = 2790$ $2790 + 40 = 2830$ $2830 + 40 = 2870$ $2870 + 40 = 2910$ $2910 + 40 = 2950$ $2950 + 40 = 2990$ $2990 + 40 = 3030$ $3030 + 40 = 3070$ $3070 + 40 = 3110$ $3110 + 40 = 3150$ $3150 + 40 = 3190$ $3190 + 40 = 3230$ $3230 + 40 = 3270$ $3270 + 40 = 3310$ $3310 + 40 = 3350$ $3350 + 40 = 3390$ $3390 + 40 = 3430$ $3430 + 40 = 3470$ $3470 + 40 = 3510$ $3510 + 40 = 3550$ </p>

Figure 7. Students' Answers to Question 4

In question 4, the cognitive domains to be achieved are C3 (applying). Based on students' answers, that the two dominant patterns of student answers are (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula, but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

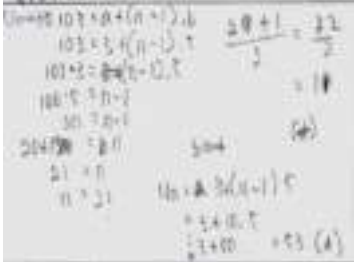

Question 5:	Answer Pattern 1	Answer Pattern 2
<p>Diketahui barisan aritmetika 3, 8, 13, 18, ..., 103. Suku tengah dari barisan tersebut adalah ...</p> <p>a. 53 b. 52 c. 20 d. 11 e. 10</p>		

Figure 8. Students' Answers to Question 5

In question 5, the cognitive domain to be achieved is C1 (remembering). Based on students' answers, that the two dominant patterns of student answers are (1) students have understood and can determine the middle term in an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use a general formula, but by writing the terms its terms from the known terms and inserts several terms and then determines them.

In addition to knowing students' understanding and mastery, interviews were conducted to see student responses to the assessment for learning instrument. Based on the results of the interview, it was found that the students thought that the assessment for learning instruments could increase learning motivation to be even better, and post-learning assessment is very important because students can know their ability to understand what is conveyed by the teacher.

Description of Student Ability

The instrument produced from this research is in the form of a multiple-choice test accompanied by open reasons and has met the established criteria, both validity, reliability, and item parameters. This instrument can explore students' mathematical thinking processes more deeply. This instrument has combined multiple-choice and essay tests. Multiple-choice tests are easier to check students' answers but students' mathematical thinking processes cannot be explored in depth. While the essay test has the advantage of being able to explore the mathematical thinking process more deeply but it takes a long time to check the answers.

An assessment for learning instrument that has done item analysis, and the results of student analysis is one of the important sources of composite scores to be reported. In the final report, the test taker's ability score must be transformed into a score of 0 - 10 or 0 - 100, according to the needs of the school. This transformation can be done using a linear transformation by dividing the acquisition score by the ideal score and then the result is multiplied by 10 to obtain a value in the range 0 – 10 or multiplied by 100 to obtain a value of 0 – 100. In the range, 0 – 10, the value obtained by students who take the assessment for learning tests in mathematics the highest is 8.56 and the lowest is 4.31. In the range of 0 – 100, the score obtained by students who take the assessment for learning tests in mathematics is the highest 85.625 and the lowest is 43.125.

The results of the assessment of students' ability in mathematics are presented in the form of very low to very high predicates using scoring categories. The results of the analysis of the assessment for learning instrument test showed that most of the students had low and very low abilities, namely 62% (253 people). Meanwhile, students who have high and very high abilities are 38% (160 people). Other analysis results found that students who have high abilities tend to work according to the concepts that have been given by the teacher but do not follow the completion steps, students who have moderate abilities can solve problems according to the concepts that have been given by the teacher and the steps, and students who have the ability Some are unable to use the concepts given by the teacher and are not even able to give clear reasons.

Another result of this research is that the mathematics teachers involved in this study stated that the teachers agreed to provide an assessment for learning with open-ended multiple-choice questions. The reason is that this assessment can make it easier for teachers to find out the teacher's difficulties in exploring students' difficulties in certain materials. In this way, the

teacher can provide remedial or other assistance to students who have learning difficulties. This means that assessment for learning with Polytomous Response can be used as a way to determine which students need to get remedial or not. Generally, previous research mentions how to determine students who need remedial only using one test, namely multiple-choice tests (Gierl, Bulut, Guo, & Zhang, 2017) or essays (Putri, Kartono, & Supriyadi, 2020)

Conclusion

Based on the results of the analysis that the researcher has done, it can be concluded that all the assessments for learning items with the response polytomous that have been developed are compatible with the Partial Credit Model (PCM). The overall assessment for learning items with response polytomous that have been developed have difficulty levels in the medium category. The assessment for learning an instrument with response polytomous that has been developed based on the results of construct validity consists of 40 items. The assessment for learning an instrument with response polytomous has been developed as an index of content validity of 1 or very high category. The assessment for learning an instrument with the response polytomous that has been developed has a reliability coefficient of 0.89 or a very high category.

Recommendations

Based on the results of the study, there are several recommendations for vocational teachers, schools, and other researchers. For teachers, before using this instrument, the teacher should familiarize students with giving questions with polytomous responses. For schools, they should encourage other teachers to take advantage of this test and recommend standardizing the tests given to students. For other researchers, this research can be used as a reference in developing test instruments for other materials.

Limitations

The research conducted has several limitations, namely the selection of schools used as samples is not by the expectations of researchers so that schools with high, medium, and low quality are not yet representative. This is due to the very long distance from one school to another. The learning process in the classroom is not fully controlled by the researcher so that the students who are the research sample are less conditioned. In addition, this research is limited to the scope of material for sequences and series, matrices, and equations, and quadratic functions for vocational schools.

References

- Arikunto, S. (2012). *Dasar-Dasar Evaluasi Pendidikan Edisi 2* [Educational Evaluation Fundamentals Edition 2]. Jakarta: Bumi Aksara. Translation not italic, in sentence case
- Baker, F. B. (2001). *The Basics of Item Response Theory*: ERIC, <https://eric.ed.gov/?id=ED458219>.
- Bloom, B. (1956). *Taxonomy of Education Objectives: Handbook 1, Cognitive Domain*. New York: David McKay.
- Buckles, S., & Siegfried, J. J. (2006). Using Multiple-Choice Questions to Evaluate in-Depth Learning of Economics. *The Journal of Economic Education*, 37(1), 48-57. <https://doi.org/10.3200/JECE.37.1.48-57>
- Çanakkale, G., & Çanakkale, G. (2013). Developing a Science Process Skills Test Regarding The 6th Graders. *The International Journal of Assessment and Evaluation*, 19, 39-57. <https://doi.org/10.18848/2327-7920/CGP/v19i02/48322> Incorrect authors. Please check the original paper and add correct authors
- Chandrasegaran, A., Treagust, D. F., & Mocerino, M. (2007). The Development of a Two-Tier Multiple-Choice Diagnostic Instrument for Evaluating Secondary School Students' Ability to Describe and Explain Chemical Reactions Using Multiple Levels of Representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F> 1) space after "&" 2) Year in parentheses
- Demars, C. (2010). *Item Response Theory*. Oxford: University Press. 1) DeMars 2) Year in parentheses 3) Add DOI link
- Ellery, K. (2008). Assessment for Learning: A Case Study Using Feedback Effectively in an Essay-Style Test. *Assessment & Evaluation in Higher Education*, 33(4), 421-429. <https://doi.org/10.1080/02602930701562981> Year in parentheses
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analysing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, 87(6), 1082-1116. <https://doi.org/10.3102/0034654317726529>

- Haladyna, T. M. 2004. *Developing and Validating Multiple-Choice Test Items*: Routledge. 1) Year in parentheses 2) Add DOI link
- Hambleton, & Jones, R. W. (1993). Comparison Of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, 12(3), 38-47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x> 1) Missing initial of the author 2) space after "&"
- Hambleton, Swaminathan, H., & Rogers, H. (1985). *Principles and Applications of Item Response Theory*. Boston, MA: Lower-Nihon Publishing Company. 1) Missing initial of the author 2) Please double-check the authors and publisher and make sure they are correct.
- Hambleton, Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory* (Vol. 2). London: Sage Publications. Missing initial of the author
- Hanafi, N. M., Ab Rahman, A., Mukhtar, M. I., Ahmad, J., & Warman, S. (2014). Validity and Reliability of Competency Assessment Implementation (CAI) Instrument Using Rasch Model. *International Journal of Psychological And Behavioral Sciences*, 8(1), 162-167. <https://doi.org/10.5281/zenodo.1336562> 1) "And" should be in lowercase 2) space after "&"
- Hirsch, L. S., & O'Donnell, A. M. (2001). Representativeness in Statistical Reasoning: Identifying and Assessing Misconceptions. *Journal of Statistics Education*, 9(2). <https://doi.org/10.1080/10691898.2001.11910655> space after "&"
- Krishnan, S. R., & Howe, A. C. (1994). The Mole Concept: Developing an Instrument to Assess Conceptual Understanding. *Journal Of Chemical Education*, 71(8), 653. <https://doi.org/10.1021/ed071p653> 1) space after "&" 2) "Of" should be in lowercase
- Malhotra, N., Nunan, D., & Birks, D. (2017). *Marketing Research: An Applied Approach*: Pearson. <http://www.pearsoned.co.uk/bookshop/detail.asp?item=100000000589380>
- Masters, G. N. (2011). The Partial Credit Model *Handbook of Polytomous Item Response Theory Models* (Pp. 119-132). New York: Routledge.
- Nitko, A. J. (1996). *Educational Assessment of Students*: ERIC. <https://eric.ed.gov/?id=ED435654>
- Ostini, R., & Nering, M. L. (2006). *Polytomous Item Response Theory Models*. London: Sage Publications. Add DOI link
- Plomp, T. (2013). *Educational Design Research: An Introduction*. Netherlands: Netherlands Institute for Curriculum Development (SLO).
- Putri, B., Kartono, & Supriyadi. (2020). Analysis of Essay Test Instruments Using Higher Thinking Skill (HOTS) at High School Mathematics Students Using The Rasch Model. *Journal of Educational Research and Evaluation*, 9(2), 58-69. <https://doi.org/10.15294/jere.v9i2.46133>.
- Retnawati, H. (2014). *Teori Respons Butir dan Penerapannya: Untuk Peneliti, Praktisi Pengukuran dan Pengujian, Mahasiswa Pascasarjana* [Item Response Theory and Its Application: For Researchers, Measurement and Testing Practitioners, Graduate Students]. Yogyakarta: Nuha Medika.
- Retnawati, H. (2016a). Proving Content Validity of Self-Regulated Learning Scale (The Comparison of Aiken Index And Expanded Gregory Index). *Reid (Research And Evaluation in Education)*, 2(2), 155-164. <https://doi.org/10.21831/reid.v2i2.11029>

- Retnawati, H. (2016b). *Validitas, Reliabilitas dan Karakteristik Butir* [Item Validity, Reliability and Characteristics]. Parama Publishing.
- Roediger III, H. L., & Marsh, E. J. (2005). The Positive and Negative Consequences Of Multiple-Choice Testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 1155. <https://doi.org/10.1037/0278-7393.31.5.1155> space after "&"
- Subali, B., & Suyata, P. (2011). *Panduan Analisis Data Pengukuran Pendidikan untuk Memperoleh Bukti Empirik Kesahihan Menggunakan Program Quest* [Education Measurement Data Analysis Guide to Obtaining Empirical Evidence of Validity Using the Quest Program]. Yogyakarta: Lembaga Penelitian dan Pengabdian Pada Masyarakat UNY.
- Suhaini, M., Ahmad, A., & Bohari, M. (2021). Assessments on Vocational Knowledge and Skills: A Content Validity. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Syaifuddin, M. (2020). Implementation of Authentic Assessment on Mathematics Teaching: Study on Junior High School Teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Torres, C. M. D. P., Lopes, A. P., Azevedo, J. M. M. L., & Babo, M. D. L. (2009). *Developing Multiple-Choice Questions in Mathematics*. <https://www.semanticscholar.org/>
- Whitehead, J. C., Huang, J.-C., Blomquist, G. C., & Ready, R. C. (1998). Construct Validity of Dichotomous and Polychotomous Choice Contingent Valuation Questions. *Environmental and Resource Economics*, 11(1), 107-116. <https://doi.org/10.1023/A:1008231430184>
- William, D. (2011). What is Assessment for Learning?. *Studies in Educational Evaluation*, 37(1), 3-14. <https://doi.org/10.1016/j.stueduc.2011.03.001>
- Williams, J. B. (2006). Assertion-Reason Multiple-Choice Testing as A Tool For Deep Learning: A Qualitative Analysis. *Assessment & Evaluation in Higher Education*, 31(3), 287-301. <https://doi.org/10.1080/02602930500352857>

SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Tue, Feb 15, 2022 at 11:48 PM

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research

Until now I am still doing the revision process, because of the many revision suggestions from reviewers. Therefore, may I be allowed to extend the revision time (approximately one week)?

Your deadline: February 15, 2022.

Thank you for your kindness.

Sugeng Sutiarso
Lampung, Indonesia.
[Quoted text hidden]

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Wed, Feb 16, 2022 at 12:27 AM

Dear Dr. Sutiarso,

Thank you for your kind reply. We have extended the deadline to February 22, 2022.

We are looking forward to getting your revised paper by February 22.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com

[Quoted text hidden]

SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Tue, Feb 22, 2022 at 8:48 PM

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.


I have revised the article according to the reviewer's suggestion. Here I attach (1) a revised article, (2) a correction report, and (3) a proofreading certificate from my university's language center.

Best regards,

Sugeng Sutiarso
Lampung University

3 attachments

 **CORRECTION REPORT_Article Sugeng Sutiarso et al.docx**
31K

 **Revision_Article Sugeng Sutiarso et al.docx**
1566K

 **Cert of Proofreading_Sugeng Sutiarso et al.pdf**
275K

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Tue, Feb 22, 2022 at 10:16 PM

Dear Dr. Sutiarso,

We have received your revised paper and correction report. We have sent them to our reviewers again in order to

CORRECTION REPORT			
No	Reviewer Code	Reviews	Corrections made by the author
1.	R2612	<p>Introduction: This study aims to develop an assessment test using polytomous item response theory. However, the introduction is overly long and confusing. There is much information about the item response theories. However, <u>most of this information should be moved to the method.</u> Instead, in the introduction, the authors <u>should focus on the strengths and weaknesses of the existing instruments</u> that developed to assess students' learning in mathematics in the context of vocational/high school (?) students.</p>	<p>Introduction: <u>Moving modern theoretical data analysis from introduction to research methods:</u> 2.2 <u>Analysis of test data with modern theory</u> a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items identifies the ability of about 50% of respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column. b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column. According to modern theory, before analyzing the level of difficulty and item discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton, et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigen values of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals on the unidimensional solution (DeMars, 2010). In this study, the dimensional test uses the Eigen value analysis of the correlation matrix between items. Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer on each item. If the unidimensional assumption has been met, the local independence assumption has also been met (DeMars, 2010). Model fit test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring Outfit Mean Square (MNSQ) and Pt-Measure. If the Outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and Standard Error Measurement (SEM) are analyzed which aims to further explain the latent ability as</p>

		<p>The current version of the introduction opens many questions. These are: <u>What do we know about the existing instruments? What do we need to know about the existing instruments?</u> Why is this study important for students and teachers, and scholars? These need to be answered in this section.</p>	<p>measured by using a test that is expressed through item donations. (pp. 9-10)</p> <p><u>Focus on the strengths and weaknesses of the existing instruments:</u> Each form has advantages or disadvantages to each other. The advantage of multiple-choice over essays is that multiple-choice can measure multiple cognitive levels, scoring more objectively, and saving time. Meanwhile, essay tests can only describe students' actual abilities, but scoring tends to be subjective. (p.2)</p> <p><u>What do we know about the existing instruments?</u> The results of the preliminary research survey found that so far teachers have used multiple choice tests to determine students' abilities. During the test, students tend to guess on difficult items, sometimes students guess right and sometimes wrong. Of course, it is difficult to distinguish between students who answered correctly and guessed. Therefore, it is necessary to develop multiple choice tests that prevent students from guessing the answers.(p. 2)</p> <p><u>We need to know about the existing instruments?</u> Related to the polytomous response test, a measurement innovation is needed that can guarantee a good test according to classical theory and modern theory. This research is development research that aims to produce a good instrument of assessment in mathematics using polytomous response according to classical and modern theories. (p.5)</p> <p><u>Why is this study important for students and teachers, and scholars?</u> If the results of this research are obtained a proper test according to classical and modern theory, it can make it easier for teachers to develop appropriate tests on other materials, and students can find out their true abilities. Finally, for practitioners or other researchers, it can be a reference in further research. (p.5)</p>
2.	R2612	<p>Method: More information regarding participants should be given. For example, we need to have information about their grade, grade level, and gender.</p> <p>How many experts were enrolled to assess the validity? Please give details. In addition, what were the</p>	<p>Method: The selected schools are three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects are 413 students at grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 from the ideal score of 100). (p. 6)</p> <p>The instrument is validated by two people who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert are the suitability of the items with the indicators, language, and alternative answers to the questions. (p. 7) In</p>

		<p>experts' feedbacks on the draft version of the instrument?</p> <p>It is not clear how the polytomous item response theory was implemented in the method and the results. More details are needed</p>	<p>addition to providing assessments, the experts also provide some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arrange them in order. (p.12)</p> <p>Method: This research is a research and development that refers to Plomp's model (Plomp, 2013) with the research procedure consisting of five stages, namely preliminary investigation, design, realization or construction, test phase, revision, and implementation (test). The preliminary investigation stage is to identify the assessment instruments used by teachers so far. The design stage is to make a polytomous response test grid according to the basic competencies of mathematical concepts and to make an expert assessment questionnaire sheet. The realization or construction stage is developing the items on the polytomous response test, and also the expert validation process for the items on the developed polytomous response test. The revision stage is the improvement of items on the polytomous response test based on expert advice. The implementation (test) stage is to try out the polytomous response test to students and analyze the results of the test. (pp.5-6)</p> <p>Result (pp. 14 – 21)</p> <p><i>1.2 Analysis of Test Data with Modern Theory</i> <i>Unidimensional Assumption Test</i></p> <p>The unidimensional assumption test is the first assumption test with factor analysis. Factor..... In conclusion, the characteristics of the test kit are suitable for students with moderate abilities.</p>
3.	R2612	<p>Results: Figure 5 is not addressed in the text.</p>	<p>Results: Figure 5 shows that the instrument provides a maximum of 22.36 information and has the smallest measurement error of 0.21 if it is given to students with moderate ability, which is 0.2. The lower limit and upper limit of the interval is the ability score where the graph of the information function and the SEM graph intersect in that interval. The graph indicates that the greater the value of the information function, the smaller the measurement error (SEM). Item information function states the strength or contribution of test items in revealing the latent trait as measured by the test. With the item information function, it is known which items match the model, thus helping in the selection of test items (Retnawati, 2014). In conclusion, the characteristics of the test kit are suitable for students with moderate abilities. (p. 21)</p>

		<p>Discussion The authors did write the following sentence: “One of the reasons that make this instrument suitable for use in the process of preparing the instrument under expert supervision.” However, I do not agree with the comment of the authors as the expert view has been received for all developed instruments.</p> <p>The discussion is overly long and hard to read and follow the text. The authors did still repeat the results in this section. This repetition of the results makes the text hard to understand and misses the focus of the study. Therefore, I suggest that the authors should make a brief overview of the results. Later, they should discuss similarities between the existing instruments in the literature and the developed instrument in this study. After this, they should discuss the differences between previous studies and this study. While discussing the similarities and differences, the author should also discuss possible reasons for the results.</p>	<p>Discussion The sentence "aaaa" has been removed, and replaced with a comparison between classical and modern theoretical analysis.</p> <p>The discussion has included (1) a brief overview of the results, (2) analysis of the similarities and differences between the previous instrument, and (3) the developed instrument, as well as possible reasoning for the future.</p> <p>(1) A brief overview of the results This research is development research to produce an instrument using polytomous response. The instrument is a multiple-choice test with open reasons. This instrument is analyzed by classical and modern theory. There are differences in the results of the analysis between classical and modern theories, namely item discrimination. Classical theory analysis obtains 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtains 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as not good but the results of the modern analysis are categorized as good, and vice versa (Retnawati, 2014). It means, if you find items that are not in a good category with classical theory, do not be in a hurry to revise or replace them before the analysis of modern theory analysis.</p> <p>(2) Analysis of the similarities and differences between the previous instrument Research on assessment for learning with polytomous responses with multiple-choice questions (having open reasons) is still limited. When compared to previous research, there is only one study on assessment for learning with polytomous responses to open-ended multiple-choice questions (Yang, et al., 2017). However, Yang study has several fundamental differences, namely the research objectives and data analysis. The research aims to diagnose student errors in university on the concept of calculus, not to produce a good assessment instrument. The data analysis uses parametric statistics (covariance), not using item analysis (classical and modern). Since the objectives and data analysis are different, the results of the study cannot be compared with the results of this study. However, this research has provided a reference for researchers in making reasoned multiple-choice questions, such</p>
--	--	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

		<p>The results of analyzing students' answers must be moved to the results. In the discussion, we need the results themselves.</p>	<p>as the suitability of items with indicators, language, and alternative answers to questions.</p> <p>Other studies are similar to assessment for learning with polytomous responses on multiple-choice questions with reasons (Sarea, 2018). The similarity of Sarea's research is safe, it lies in the research objectives and the analysis used (classical and modern). However, the difference is the researchers do not develop their questions and the questions are in the form of closed multiple-choice questions. The results of Sarea's research states that the comparison of the results of the classical and modern methods of item analysis is different. The difference is that the level of difficulty and item discrimination in the classical method is more categorized as good than the modern method. In other words, the modern way of stating the level of difficulty and item discrimination is categorized as good even though the analysis method states that the items are categorized as not good. Likewise with Saepuzaman's research, namely the product developed in the form of multiple-choice questions with closed reasons. Although the products are different, the results of the research can contribute to this research. The contribution of both is that there is a belief in the items that are declared not good by classical analysis, which turn out to be good with modern analysis.</p> <p>(3) The developed instrument, as well as possible reasoning for the future.</p> <p>The results of previous studies have provided support to the results of the research that the polytomous response instrument in the form of multiple-choice questions with open reasons can be used as an alternative assessment for learning, as well as other assessments (assessment as learning and assessment of learning) for all vocational schools in Lampung, Indonesia and even outside Indonesia.</p> <p>The results of the analysis of student answers have been moved to results.</p> <p>Based on the test results, in addition to knowing the quality of the developed test instruments, it can also be seen the ability of students to work on the questions given. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of 6 student answers were selected as samples with different abilities (high, medium, and low).</p> <p>Item 1: the cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms. arithmetic sequence..... and so on (pp. 21-24)</p>
--	--	------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

4.	R2612	<p>Conclusion: What is the new knowledge from this research for researchers and the literature? Please explain this detail.</p>	<p>Conclusion: It means if you find items that are not in a good category with classical theory, do not be in a hurry to revise or replace them before the analysis of modern theory analysis. The condition that must exist so that an instrument with a response polytomus can be used to determine students' abilities is the suitability of the results of the analysis between classical and modern theories. (p.27).</p>
5.	R2612	<p>Recommendations: Please make more specific suggestions for research in the future.</p>	<p>Recommendations: In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomus (pretest). This is important, so that students' prior knowledge can be known so that learning can be effective.</p>
6.	R2612	<p>Language The use of the language is very problematic. The text needs proofreading by a native speaker.</p>	<p>Language: The text has been proofread by the Language Center of the University of Lampung (the proof of certificate is attached in another file).</p>
7.	R2613	<p>Discussion section seems like more detailed findings section. Use previous studies to compare or contrast your results in Discussion section.</p>	<p>Discussion: Research on assessment for learning with polytomous responses with multiple-choice questions (having open reasons) is still limited. When compared to previous research, there is only one study on assessment for learning with polytomous responses to open-ended multiple-choice questions (Yang, et al., 2017). However, Yang study has several fundamental differences, namely the research objectives and data analysis. The research aims to diagnose student errors in university on the concept of calculus, not to produce a good assessment instrument. The data analysis uses parametric statistics (covariance), not using item analysis (classical and modern). Since the objectives and data analysis are different, the results of the study cannot be compared with the results of this study. However, this research has provided a reference for researchers in making reasoned multiple-choice questions, such as the suitability of items with indicators, language, and alternative answers to questions.</p> <p>Other studies are similar to assessment for learning with polytomous responses on multiple-choice questions with reasons (Sarea, 2018). The similarity of Sarea's research is safe, it lies in the research objectives and the analysis used (classical and modern). However, the difference is the researchers do not develop their questions and the questions are in the form of closed multiple-choice questions. The results of Sarea's research states that the comparison of the results of the classical and modern methods of item analysis is different. The difference is that the level of difficulty and item discrimination in the classical method is more categorized as good than the modern method. In other words, the modern way of stating the level of difficulty and item discrimination is categorized as good even though the analysis method states that the items are categorized as not good. Likewise with</p>

			<p>Saepuzaman's research, namely the product developed in the form of multiple-choice questions with closed reasons. Although the products are different, the results of the research can contribute to this research. The contribution of both is that there is a belief in the items that are declared not good by classical analysis, which turn out to be good with modern analysis. The results of previous studies have provided support to the results of the research that the polytomous response instrument in the form of multiple-choice questions with open reasons can be used as an alternative assessment for learning, as well as other assessments (assessment as learning and assessment of learning) for all vocational schools in Lampung, Indonesia and even outside Indonesia. (pp.26-27)</p>
8.	R2613	Add recommendations for future research to Recommendations section.	<p>Recommendations: In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomus (pretest). This is important, so that students' prior knowledge can be known so that learning can be effective.</p>

DEVELOPMENT OF ASSESSMENT FOR LEARNING INSTRUMENT USING POLYTOMOUS RESPONSE IN MATHEMATICS IN VOCATIONAL SCHOOLS

Abstract: This research is development research that aims to produce a good instrument of assessment in mathematics using polytomous response according to classical and modern theories. Plomp's model as a research design has five stages: preliminary investigation, design, realization or construction, trial, revision, and implementation (testing). The study is conducted in three vocational schools in Lampung, Indonesia. The study involves 413 students consisting of 191 male students and 222 female students. The data are collected through questionnaires and tests. Questionnaire is to identify instruments commonly used by teachers so far and to validate instruments by experts. The test uses multiple-choice tests with open reasons as many as 40 items. The data are analyzed in two ways, namely analysis with classical and modern theories. The results show that the instrument of assessment in mathematics with polytomous response has good categories according to classical and modern theory, although item discrimination according to classical theory needs to be revised, and the instrument of assessment in mathematics with the polytomous response (multiple-choice tests with open reason) can provide information on the actual competency of students. It is evidenced by the suitability of the results of the analysis of classical and modern theories.

Key words: assessment for learning, classical and modern theory, multiple choice tests with open reason, polytomous response, vocational school

Introduction

Assessment is an important activity and needs to be done by teachers in schools. The conventional paradigm often interprets the assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syarifuddin, 2020). If referring to the current paradigm, assessment in schools is divided into three parts, namely assessment as learning, assessment

for learning, dan assessment of learning (Wulan, 2018). Assessment as learning has almost the same function as assessment for learning but assessment as learning involves students in assessment, such as assessing themselves or colleagues. Assessment of learning is an assessment carried out after all learning ends and aims to assess student achievement. Assessment for learning is an assessment carried out during the learning process and aims to improve learning. It can play a role in preventing students from experiencing further learning failure because of its position between the other two assessments (Earl, 2013) as shown in the following assessment pyramid (Figure 1).



Figur 1. Assesment pyramid

Assessment activities can be applied with tests. A test is a tool or procedure used to find out or measure students' abilities about something with certain rules (Arikunto, 2012). According to the type, the test consists of two types, namely multiple choice and essay. Each form has advantages or disadvantages to each other. The advantage of multiple-choice over essays is that multiple-choice can measure multiple cognitive levels, scoring more objectively, and saving time. Meanwhile, essay tests can only describe students' actual abilities, but scoring tends to be subjective (Rosidin, 2017). The results of the preliminary research survey found that so far teachers have used multiple choice tests to determine students' abilities. During the test, students tend to guess on difficult items, sometimes students guess right and sometimes wrong. Of course, it is difficult to distinguish between students who answered correctly and

guessed. Therefore, it is necessary to develop multiple choice tests that prevent students from guessing the answers.

Specifically for choice tests, in the last four decades, evaluation experts have developed multiple-choice tests, namely multiple-choice tests with open and closed reasons (Suwarto, 2012). Multiple choice test with open reasons is a multiple-choice test that has alternative answers and questions are asked for the choice of choice. Meanwhile, multiple-choice tests with closed reasons are multiple-choice tests that have alternative answers, and students are asked to choose the reasons provided. The advantage of the choice test, compared to the free choice is the answer from the answer chosen, compared to the multiple tests which are not free to give reasons. Generally, multiple-choice tests on assessment for learning do not provide reasons (open or ask closed). The multiple-choice test has only true or false answers. If the answer is correct, then it gets a score of 1. Otherwise, if the answer is wrong, it gets a score of 0, and this scoring is called dichotomous. The choice test provides a choice of opportunities for students to get a score even though the answer is wrong and this scoring is called polytomous (Kartono, 2008).

The first time, the polytomous response test was in the 80s, and the test was known as the two-tier multiple-choice test (Treagust, 1988). The test is in the form of multiple-choice with closed reasons which aim to diagnose errors in the concepts of biology, physics, and chemistry. Multiple choice test with closed reasons has two levels, the first level is a multiple-choice test and the second level is a multiple-choice test with a choice of reasons from the first level questions (Chandrasegaran, et al., 2007). Furthermore, studies on the development of multiple-choice tests with closed reasons have been carried out by many researchers on several concepts, such as physics on light and optical instruments (Widiyatmoko & Shimizu, 2018), mathematics on calculus (Khiyarunnisa & Retnawati,

2018), chemistry on acids and base (Andaria & Hadiwinarto, 2020), biology on the respiratory system (Myanda, et al., 2020), mathematics on reasoning problems (Ambarwati, et al., 2020), biology on the human digestive system (Jamhari, 2021), mathematics on mathematical connections (Lestari, et al., 2021). Meanwhile, the development of multiple-choice tests with open reasons is still limited, namely mathematics in calculus (Yang, et al., 2017), and outside mathematics, namely physics in Higher Order Thinking Skills (Prasetya, et al., 2019).

There are two approaches to analyzing test items, namely classical and modern theory. Classical theory has weaknesses. It cannot separate the characteristics of respondents and test items. It means that the test taker's ability is only determined by the test. The characteristics of the test items will change when the examinees change, and the characteristics of the examinees will change when the items change. Thus, classical theory is considered less able to provide information on the actual abilities of students because the results of the assessment are highly dependent on the respondents. Modern theory, with item response theory, is a solution to overcome the weakness of classical theory because item response theory has the concept of releasing the relationship between respondents and test items (Saepuzaman, et al., 2021). It means a test that has a good category not only according to classical theory but must be supported by test analysis according to modern theory.

Classical theory is a measurement theory for assessing tests based on the assumption of measurement errors between actual results and observations, or correlation measurement errors based on test-takers. From the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a theory of measurement to assess by comparing the average performance against the appearance of evidence of group ability, or the ability of test-takers predicted from their

abilities; or better known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical test theory is used because it is easy to apply but has limitations in measuring the level of difficulty and differentiating items because the calculation of the two indicators is based directly on the total score of the test takers. While modern theory with item response theory frees the dependence between test items and respondents (parameter invariance concept), responses to one test item do not affect other test items (local independence concept), and test items only measure one dimension or unidimensional concept (Anisa, 2013).

Related to the polytomous response test, a measurement innovation is needed that can guarantee a good test according to classical theory and modern theory. This research is development research that aims to produce a good instrument of assessment in mathematics using polytomous response according to classical and modern theories. If the results of this research are obtained a proper test according to classical and modern theory, it can make it easier for teachers to develop appropriate tests on other materials, and students can find out their true abilities. Finally, for practitioners or other researchers, it can be a reference in further research. The formulation of the problem proposed is (1) does the instrument in the assessment in mathematics developed with the polytomous response have a good test category according to classical and modern theory measurements?, and (2) does the assessment for learning an instrument with the polytomous response provide information on the actual competence of students?

Research Method

Research Design

This research is a research and development that refers to Plomp's model (Plomp, 2013) with the research procedure consisting of five stages, namely preliminary investigation, design,

realization or construction, test phase, revision, and implementation (test).

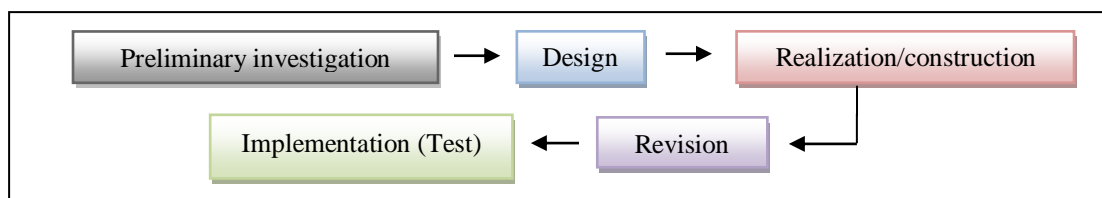


Figure 2. Stages of research design

The preliminary investigation stage is to identify the assessment instruments used by teachers so far. The design stage is to make a polytomous response test grid according to the basic competencies of mathematical concepts and to make an expert assessment questionnaire sheet. The realization or construction stage is developing the items on the polytomous response test, and also the expert validation process for the items on the developed polytomous response test. The revision stage is the improvement of items on the polytomous response test based on expert advice. The implementation (test) stage is to try out the polytomous response test to students and analyze the results of the test.

Research Subject

The subjects of the study are students of a vocational school in the province of Lampung, Indonesia. The research sample is determined using a non-probability sampling technique in the form of accidental sampling. It means taking the subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools are three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects are 413 students at grade I (male students = 191, and female students = 222), whose mathematical abilities on the

National Exam (NE) were categorized as moderate (average 64.67 from the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
SMK Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collecting Technique

Data are collected using a questionnaire and test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed and to determine content validity (Suhaini, et al., 2021). The instrument is validated by two people who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert are the suitability of the items with the indicators, language, and alternative answers to the questions. The score in validating the instrument follows the criteria as in Table 2 below.

Table 2. Content validity score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining content validity, the instrument is tested on students. Then, it is continued by determining the validity of the construct and its reliability, with the aim that the instrument can be further analyzed.

The instrument used is a multiple choice test with a polytomus response with open reasoning, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contains five answer choices along with the reasons. Student answers score refers to the polytomus score in the Partial Credit Model, where answer choices and reasons are related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Student answers score

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The research data obtained are analyzed in two stages, namely (1) questionnaire data analysis (qualitative analysis), and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis.

1. Questionnaire data analysis (qualitative analysis)

The questionnaire data analyzed include two parts, namely identification data on the instruments used by the teacher and expert validation data. The identification data on the instrument are analyzed descriptively, and the expert validation data are analyzed for trends or expert agreement using the Gregory formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range 0 - 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the value of the validity of an item.

2. *Test data analysis (empirical analysis)*

After the content validity test, it is followed by construct validity and reliability tests. The construct validation test uses exploratory factor analysis. The instrument is considered to have a good construct if the explained Kaiser-Meyer-Olkin (KMO) value is more than 0.5 (Retnawati, 2014). Reliability test using Cronbach's Alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further tests can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is because they are both preliminary analyzes of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the IteMan program is used for classical theory, and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program is used because it has several advantages (Untary, et al., 2020), namely, it can analyze polytomous data and can analyze the maximum likelihood model using a 1-parameter logistic model.

2.1 *Analysis of test data with classical theory*

- a. The item difficulty level is the percentage of the number of students who answer the questions correctly or incorrectly. The difficulty of a good (medium) item is if the index is in the range of 0.3 to 0.7. If the index is below 0.3 then the item is difficult, and vice versa if the index is above 0.7 then the item is easy.
- b. Item discrimination is the item's ability to distinguish high-ability students from stupid, low-ability students. Good item discrimination if it has an index above 0.3; and if the item discrimination index is below 0.3 then the item needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items identifies the ability of about 50% of respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column.

According to modern theory, before analyzing the level of difficulty and item discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton, et al., 1991). Unidimensional means that each test item only measures one

ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigen values of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals on the unidimensional solution (DeMars, 2010). In this study, the dimensional test uses the Eigen value analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer on each item. If the unidimensional assumption has been met, the local independence assumption has also been met (DeMars, 2010). Model fit test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring Outfit Mean Square (MNSQ) and Pt-Measure. If the Outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and Standard Error Measurement (SEM) are analyzed which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

1. Analysis of Questionnaire Data

Based on the results of the questionnaire, it is found that so far the teacher has never used the polytomous response instrument with a multiple-choice test with open reasons. As many as 80% of teachers use essay tests and 20% of teachers use multiple-choice, with each instrument consisting of 2-5 items. In addition, about 10% of teachers use this assessment as a learning improvement, such as improving lesson plans and teaching models/methods. The results of the questionnaire stated that 90% of teachers who do not use assessment as an improvement in learning are caused by several aspects, such as teachers are not

understanding assessment (20%), teachers are not knowing how to analyze assessments (50%), and teachers are not knowing how to develop good assessment questions (30%). The following is a summary of the questionnaire data from the identification of assessment for learning instruments.

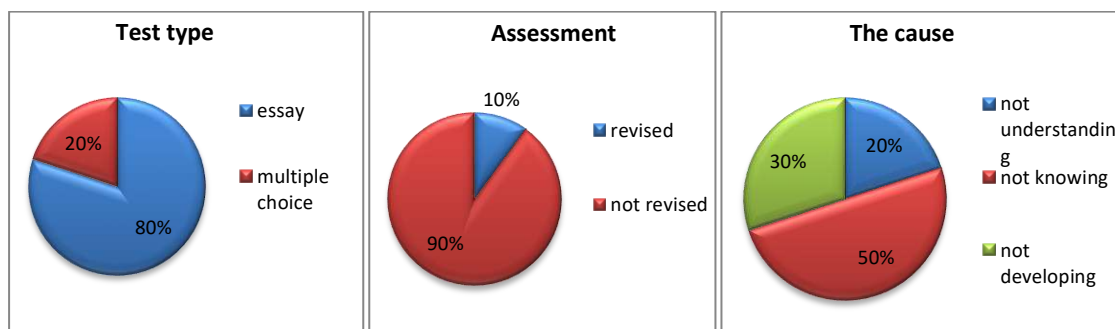


Figure 3. Description of teacher condition in assessment for learning

Content Validity

The results of the two expert assessments show that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement with the Gregory Index formula is obtained as shown in Table 4 below.

Table 4. Index Gregory items

		Rater 1	
		weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provide some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices

that are misleading, and arrange them in order.

2. Analysis of Test Data

Construct Validity

After testing the instrument, it is followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory factor analysis

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained the Cronbach's Alpha coefficient value of 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to the classical and modern methods.

Table 6. Item reliability

Cronbach's Alpha	N of Items
0.892	40

2.1 Analysis of Test Data with Classical Theory

Analysis of test data with classical theory does not require testing assumptions, but the analysis of the level of difficulty and item discrimination can be directly calculated. The results of the analysis of the level of difficulty and item discrimination are obtained as shown in Table 7 below.

Table 7. Level of difficulty and item discrimination

Item	Level of difficulty	Category	Discrimination	Category	Item	Level of difficulty	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised
4	0.540	good	0.304	Good	24	0.438	good	-0.071	revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	revised

Based on Table 7, it is found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are included in the good category. Meanwhile, only two items on discrimination have good categories, and the remaining items need to be revised. The results indicate that all items are good based on the level of difficulty, but almost all items need to be revised for item discrimination.

2.2 Analysis of Test Data with Modern Theory

Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, then calculating the eigenvalues. The eigenvalue is then used to calculate the percentage of explained variance, as well as describe the scree plot (Retnawati, 2014). The output of factor analysis is the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. KMO test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
--------------------------------------------------	------

Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

The unidimensional test is based on the cumulative percentage of eigenvalues and scree plots. If the cumulative percentage of eigenvalues in the first factor is more than 20%, then the unidimensional assumption is accepted (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the eigenvalues in the first factor is 20.220%. The cumulative percentage of the eigenvalues has exceeded 20%, so the instrument in the study is proven to only measure one factor or dimension.

Table 9. Total variance explained

Component	Initial Eigenvalues		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot which is based on the number of factors marked by the steepness of the graph with the acquisition of eigenvalues.

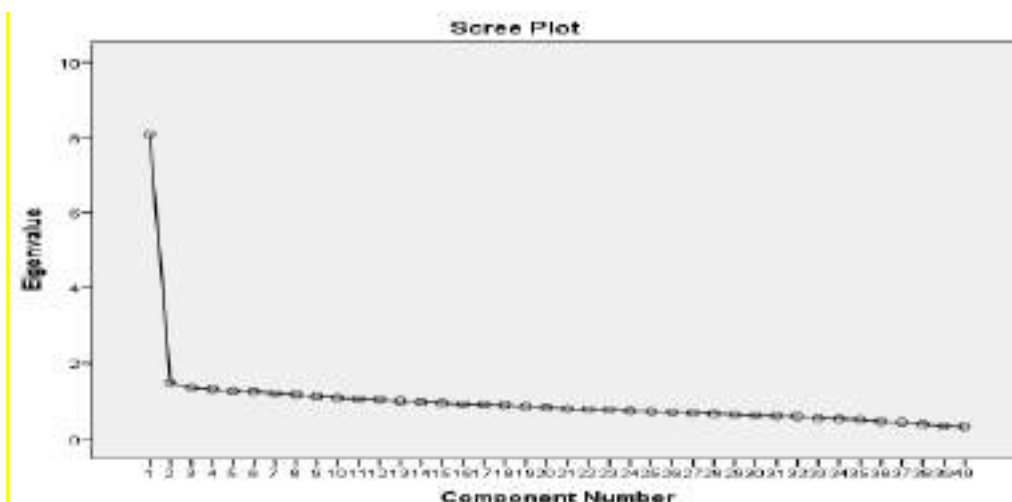


Figure 4. Scree plot unidimensi

Based on the scree plot, it is known that the eigenvalues immediately slope on the second

factor. It shows that there is only one dominant factor in the developed instrument. The results prove that the test kit meets the unidimensional assumption or in other words only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be accepted if the respondent's answer to one item does not affect the respondent's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. It confirms that the assumption is automatically proven after being proven by the unidimensionality of the respondent's data on a test (Retnawati, 2014).

Table 10. Covariance matrix

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Data in Table 10 shows the results of the variance-covariance values between groups of students' abilities. It can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. It can be concluded that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test is analyzed using the Winsteps program. The item requirements are called fit to the model if the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Corr is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is

accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items match the model or fit (Table 11).

Table 11. Item fit on model

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73
 Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Difficulty Level

The item difficulty level is analyzed using the Winsteps program, and the results obtained can be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of -2 and 2, it can be concluded that all items are in the good category. If further

divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarmo, 2015). It can also be seen on the difficult map items, namely the difficulty level is in the range of -2 and 2.

Table 12. Item difficulty level

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73
 Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .05

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ ZSTD	OUTFIT MNSQ ZSTD	PT-MEASURE CORR. EXP.	EXACT MATCH OBSN EXPN	Item
1	1134	413	-.70	.07	1.16 2.5	1.16 2.5	.74 .43	35.1 47.9	Q1
2	1079	413	-.45	.07	.97 -.4	.98 -.3	.75 .43	49.2 49.1	Q2
3	1066	413	-.39	.07	.93 -1.0	.94 -1.0	.67 .43	49.4 49.5	Q3
4	1064	413	-.38	.07	.82 -3.0	.82 -3.0	.46 .43	50.1 48.0	Q4
5	1021	413	-.19	.07	.84 -2.6	.84 -2.6	.41 .44	53.5 50.6	Q5
6	998	413	-.09	.07	1.16 2.4	1.16 2.4	.44 .44	42.4 50.9	Q6
7	977	413	.01	.07	1.03 -.3	1.03 -.5	.45 .44	40.7 51.1	Q7
8	966	413	.06	.07	1.16 2.4	1.16 2.5	.51 .44	42.1 51.1	Q8
9	965	413	.06	.07	.97 -.5	.97 -.5	.43 .44	49.9 51.0	Q9
10	964	413	.07	.07	.93 -1.0	.93 -1.0	.40 .44	49.2 51.0	Q10
11	986	413	-.03	.07	.97 -.4	.98 -.4	.38 .44	44.1 51.0	Q11
12	988	413	-.04	.07	.90 -1.7	.90 -1.7	.39 .44	52.3 51.0	Q12
13	957	413	-.10	.07	1.07 1.1	1.07 1.1	.44 .44	47.5 51.0	Q13
14	1003	413	-.11	.07	1.01 .2	1.01 .3	.42 .44	48.7 50.9	Q14
15	1000	413	-.09	.07	1.10 1.5	1.10 1.6	.49 .44	42.9 50.9	Q15
16	994	413	-.07	.07	1.08 1.2	1.08 1.2	.44 .44	46.5 51.0	Q16
17	964	413	.07	.07	.96 -.7	.96 -.6	.48 .44	52.8 51.0	Q17
18	972	413	.03	.07	1.12 1.8	1.12 1.8	.35 .44	43.1 51.1	Q18
19	974	413	.02	.07	1.04 -.7	1.05 -.8	.26 .44	47.7 51.1	Q19
20	997	413	-.08	.07	.97 -.4	.97 -.4	.47 .44	50.8 50.9	Q20
21	1010	413	-.14	.07	1.01 -.3	1.02 -.3	.44 .44	47.5 50.8	Q21
22	1050	413	-.32	.07	1.08 1.2	1.08 1.2	.58 .44	43.6 50.0	Q22
23	1037	413	-.26	.07	1.05 .8	1.05 .8	.53 .44	47.2 50.3	Q23
24	974	413	.02	.07	1.06 1.0	1.07 1.1	.37 .44	46.7 51.1	Q24
25	968	413	.05	.07	1.00 -.0	1.00 -.0	.41 .44	47.9 51.1	Q25
26	943	413	-.14	.07	.87 -.4	.87 -.4	.31 .44	50.8 50.9	Q26
27	936	413	-.20	.07	1.02 -.3	1.02 -.4	.32 .44	46.0 50.8	Q27
28	975	413	.02	.07	.89 -1.8	.88 -1.9	.40 .44	54.5 51.1	Q28
29	981	413	-.01	.07	.97 -.5	.97 -.5	.33 .44	49.2 51.0	Q29
30	949	413	-.14	.07	1.06 -.9	1.05 -.9	.30 .44	45.5 50.9	Q30
31	951	413	-.11	.07	1.07 1.1	1.07 1.1	.25 .44	44.8 51.0	Q31
32	974	413	.02	.07	1.12 1.8	1.12 1.8	.35 .44	43.3 51.1	Q32
33	981	413	-.01	.07	1.04 -.6	1.04 -.6	.41 .44	48.2 51.0	Q33
34	950	413	-.09	.07	1.06 1.0	1.07 1.1	.33 .44	46.5 51.0	Q34
35	951	413	-.11	.07	1.04 -.7	1.04 -.7	.40 .44	47.2 51.0	Q35
36	961	413	-.05	.07	1.04 -.6	1.05 -.8	.34 .44	47.7 51.0	Q36
37	955	413	-.11	.07	.96 -.6	.96 -.6	.40 .44	51.1 51.0	Q37
38	921	413	-.27	.07	.89 -1.7	.89 -1.9	.43 .44	52.8 50.6	Q38
39	825	413	-.72	.07	.65 -6.3	.66 -6.2	.19 .43	18.1 48.6	Q39
40	801	413	-.84	.07	.74 -4.5	.74 -4.6	.34 .43	15.4 48.2	Q40
MEAN	979.5	413.0	-.00	.07	1.00 -.1	1.00 -.1		47.8 50.6	
S.D.	56.8	.0	.26	.00	.11 1.8	.11 1.8		4.3 .8	

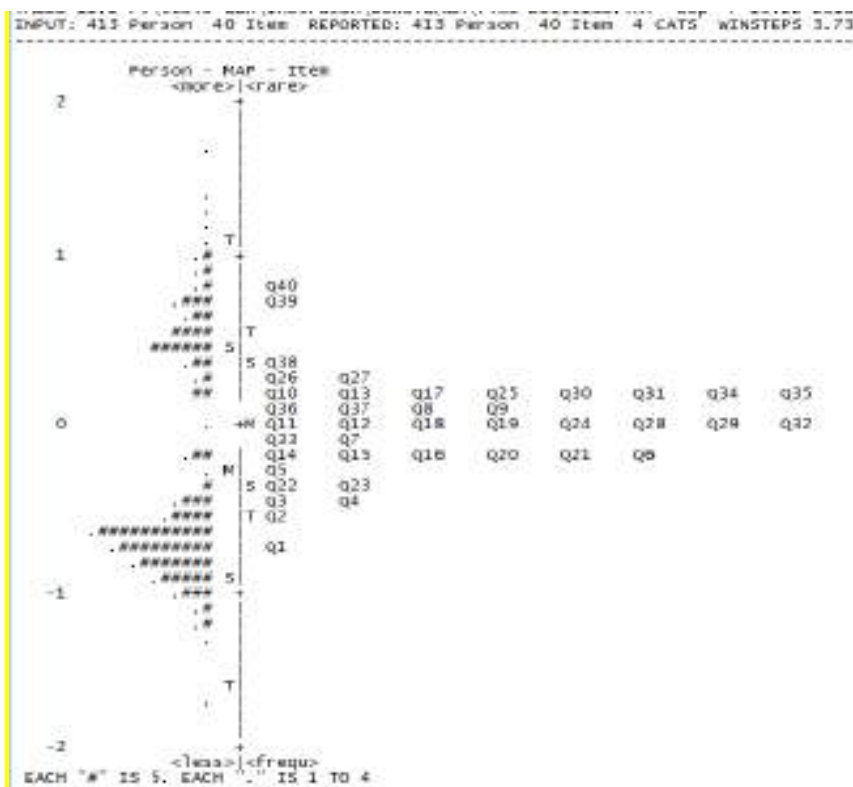


Figure 5. Item difficult map

Item Discrimination

The item discrimination analysis used the Winsteps program and the results are presented in Table 13 (column PT-Measure Corr). Item discrimination is in the range of 0.23 to 0.74; with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai, et al., 2005).

Table 12. Item discrimination

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73
 Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Comparative Analysis between Classical and Modern Theory

The results of the analysis of classical and modern theory obtained the index of difficulty level and item discrimination as follows.

Table 14. Analysis classical and modern theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Category	Percentage	Many Items with Good Category	Percentage
difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 14, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in good category than using classical theory. If we compare the index of

discriminatory items between classical (Table 7) and modern (Table 13), it can be seen that there is a match between the categories of item discrimination. It means that if the item discrimination is not good with the classical theory, then the item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability is measured by using a test that is expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error or Standard Error Measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

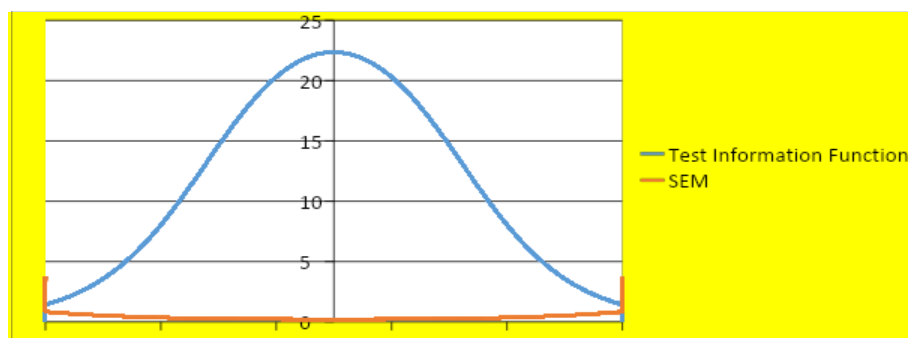


Figure 5. Graph of information function and measurement error

Figure 5 shows that the instrument provides a maximum of 22.36 information and has the smallest measurement error of 0.21 if it is given to students with moderate ability, which is 0.2. The lower limit and upper limit of the interval is the ability score where the graph of the information function and the SEM graph intersect in that interval. The graph indicates that

the greater the value of the information function, the smaller the measurement error (SEM). Item information function states the strength or contribution of test items in revealing the latent trait as measured by the test. With the item information function, it is known which items match the model, thus helping in the selection of test items (Retnawati, 2014). In conclusion, the characteristics of the test kit are suitable for students with moderate abilities.

Based on the test results, in addition to knowing the quality of the developed test instruments, it can also be seen the ability of students to work on the questions given. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of 6 student answers were selected as samples with different abilities (high, medium, and low).

Item 1: the cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms. arithmetic sequence.

Question 1:	Pattern 1:	Pattern 2:
<p>Diketahui barisan bilangan 3, 8, 13, 18,.... Rumus suku ke-n barisan itu adalah....</p> <p>a. $U_n = 5n - 3$ b. $U_n = 5n - 2$ c. $U_n = 2n + 1$ d. $U_n = 4n - 1$ e. $U_n = 3n + 2$</p> <p>Alasan:</p>	<p>1. diketahui $a = 3$ $b = 5$ ditanya $U_n = ?$ Jawab $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	<p>$U_1 = a = 3$ $b = U_2 - U_1 = 8 - 3 = 5$ $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>

Figure 6. Student answers in item 1

Item 2: the cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand and determine the number of terms in a sequence by using the general formula for an arithmetic sequence or determining the number of arithmetic sequences without using a general formula

(only writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

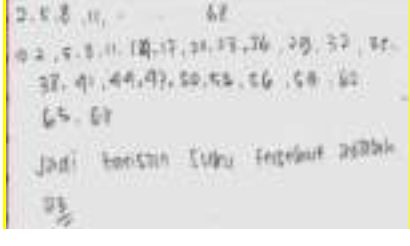
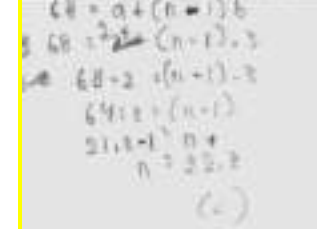
Question2:	Pattern 1	Pattern 2
<p>Diketahui barisan aritmetika 2, 5, 8, 11, ... 68. Banyaknya suku barisan tersebut adalah....</p> <p>a. 12 b. 13 c. 22 d. 23 e. 24</p> <p>Alasan:</p>	 <p>2, 5, 8, 11, ..., 68 02, 05, 08, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 59, 62 65, 68 Jadi banyak suku tersebut adalah 23</p>	 <p>$68 = a + (n-1)b$ $68 = 2 + (n-1) \cdot 3$ $66 = 3(n-1)$ $22 = n-1$ $22+1 = n$ $n = 23$</p>

Figure 7. Student answers in item 2

Item 3: the cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and determined the difference or difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

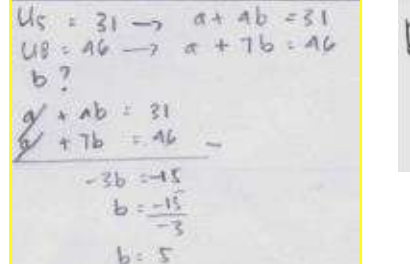
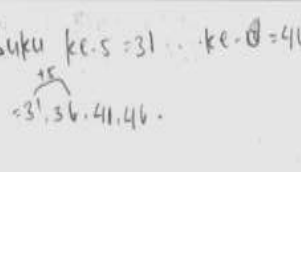
Question 3:	Pattern 1	Pattern 2
<p>Pada suatu barisan aritmetika diketahui suku ke-5 adalah 31, dan suku ke-8 adalah 46. Nilai beda barisan tersebut adalah</p> <p>a. 5 b. 6 c. 7 d. 8 e. 11</p> <p>Alasan:</p>	 <p>$U_5 = 31 \rightarrow a + 4b = 31$ $U_8 = 46 \rightarrow a + 7b = 46$ $b = ?$ $a + 4b = 31$ $a + 7b = 46$ <hr/>$-3b = -15$ $b = \frac{-15}{-3}$ $b = 5$</p>	 <p>Suku ke-5 = 31 .. suku ke-8 = 46 $\frac{46-31}{3} = 5$</p>

Figure 8. Student answers in item 3

Item 4: the cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the

general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula, but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

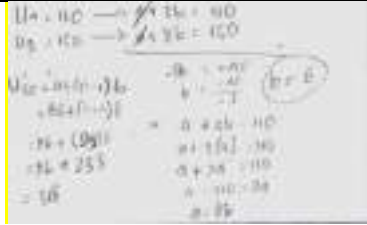
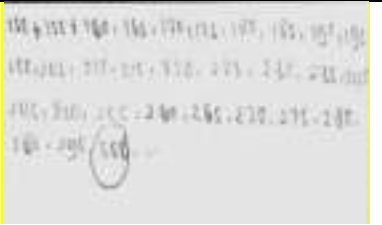
Question 4:	Pattern 1	Pattern 2
<p>Jika barisan aritmetika berturut-turut adalah 110 dan 150. Suku ke-30 barisan aritmetika tersebut adalah ...</p> <p>a. 308 b. 318 c. 326 d. 344 e. 354</p> <p>Alasan:</p>		

Figure 9. Student answers in item 4

Item 5: the cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but by writing the terms from known terms and inserts several terms and then defines them.

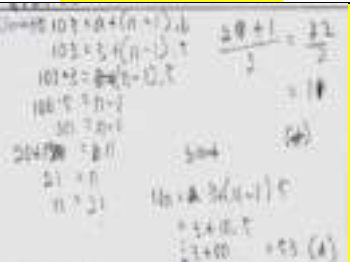

Question 5:	Answer Pattern 1	Answer Pattern 2
<p>Jika barisan aritmetika 3, 8, 13, 18, ..., 103. Suku tengah dari barisan tersebut adalah ...</p> <p>a. 53 b. 52 c. 20 d. 11 e. 10</p> <p>Alasan:</p>		

Figure 10. Student answers in item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is a multiple-choice test with open reasons and all parameters have been accepted. This instrument is a combination of multiple-choice test and essay. Multiple-choice tests are easier to check students' answers but students'

mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find out deeper mathematical thinking processes but it takes a long time to check the answers.

Assessment of learning instruments that have been carried out with item analysis and the results of student analysis is one of the important sources of composite scores to be reported. In the final report, the test taker's ability score should be changed to a score of 0 - 10 from 0 - 100, according to the needs of the school. The transformation uses a linear transformation by dividing the score by the ideal score and then the result is multiplied by 10 to get a value in the range 0 – 10 or multiplied by 100 to get a score of 0 – 100. In the range 0 – 10, the score obtained by students taking the test the highest mathematics learning was 8.56 and the lowest was 4.31. In the range 0 – 100, the scores obtained by students with the highest mathematics is 85.625 and the lowest is 43.125.

The results of the assessment of students' mathematical abilities are presented in the form of very low to very high predicates. The results of the test analysis show that most students have low and very low abilities, namely 62% (253 students). Meanwhile, students who have high and very high abilities are 38% (160 students). The results of another analysis find that students who have high abilities tend to work according to the concepts that have been given by the teacher but do not follow the completion steps, students who have moderate abilities can solve problems according to the concepts that have been given by the teacher and the steps, and there are students who have abilities but are not able to use the concepts given by the teacher and are not even able to give clear reasons..

Another result of this study is that teachers agree to provide learning assessments with multiple choice questions with open-ended reasons because the instrument is easier for

teachers to find out students' difficulties in certain materials. In this way, teachers can also provide remedial or other assistance to students who have learning difficulties. It means that the polytomous response instrument can be used as a way to determine which students need remedial or not. In general, previous research states how to determine students who need remedial only one test, namely multiple-choice tests (Gierl, et al., 2017) or essays (Putri, et al., 2020).

Discussion

This research is development research to produce an instrument using polytomous response. The instrument is a multiple-choice test with open reasons. This instrument is analyzed by classical and modern theory. There are differences in the results of the analysis between classical and modern theories, namely item discrimination. Classical theory analysis obtains 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtains 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as not good but the results of the modern analysis are categorized as good, and vice versa (Retnawati, 2014). It means, if you find items that are not in a good category with classical theory, do not be in a hurry to revise or replace them before the analysis of modern theory analysis.

Research on assessment for learning with polytomous responses with multiple-choice questions (having open reasons) is still limited. When compared to previous research, there is only one study on assessment for learning with polytomous responses to open-ended multiple-choice questions (Yang, et al., 2017). However, Yang study has several fundamental differences, namely the research objectives and data analysis. The research aims to diagnose student errors in university on the concept of calculus, not to produce a good assessment

instrument. The data analysis uses parametric statistics (covariance), not using item analysis (classical and modern). Since the objectives and data analysis are different, the results of the study cannot be compared with the results of this study. However, this research has provided a reference for researchers in making reasoned multiple-choice questions, such as the suitability of items with indicators, language, and alternative answers to questions.

Other studies are similar to assessment for learning with polytomous responses on multiple-choice questions with reasons (Sarea, 2018). The similarity of Sarea's research is safe, it lies in the research objectives and the analysis used (classical and modern). However, the difference is the researchers do not develop their questions and the questions are in the form of closed multiple-choice questions. The results of Sarea's research states that the comparison of the results of the classical and modern methods of item analysis is different. The difference is that the level of difficulty and item discrimination in the classical method is more categorized as good than the modern method. In other words, the modern way of stating the level of difficulty and item discrimination is categorized as good even though the analysis method states that the items are categorized as not good. Likewise with Saepuzaman's research, namely the product developed in the form of multiple-choice questions with closed reasons. Although the products are different, the results of the research can contribute to this research. The contribution of both is that there is a belief in the items that are declared not good by classical analysis, which turn out to be good with modern analysis. The results of previous studies have provided support to the results of the research that the polytomous response instrument in the form of multiple-choice questions with open reasons can be used as an alternative assessment for learning, as well as other assessments (assessment as learning and assessment of learning) for all vocational schools in Lampung, Indonesia and even outside Indonesia.

Conclusion

Based on the results of the research and discussion, conclusions are obtained, namely (1) the instrument with a polytomous response has been accepted according to classical and modern theory, although item discrimination in the classical theory needs to be revised; **It means if you find items that are not in a good category with classical theory, do not be in a hurry to revise or replace them before the analysis of modern theory analysis**, and (2) an instrument with a polytomous response (multiple-choice test with open reasons) can provide information on actual student competencies. It is evidenced by the suitability of the results of classical and modern analysis. **So, the condition that must exist so that an instrument with a response polytomus can be used to determine students' abilities is the suitability of the results of the analysis between classical and modern theories.**

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, teachers should familiarize students with giving a test in the form of a polytomous response before giving the test. For schools, principals or other leaders should encourage other teachers to take advantage of this test, and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. **In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomus (pretest). This is important, so that students' prior knowledge can be known so that learning can be effective.**

Limitations

The research carried out has several limitations. Firstly, the selected schools have not been the researchers' expectations, for example representing schools with high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic material (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Dordrech: Springer-Kluwer Academic Publishers.
- Ambarwati, R., Sunardi, Yudianto, E., Murtikusuma, R. P., & Safrida, L. N. (2020). Developing mathematical reasoning problems type two-tier multiple choice for junior high school students based one thnomathematics of jember fashion carnival. *ICOLSTEM*. Jember: IOP Publishing. <https://doi.org/10.1088/1742-6596/1563/1/012036>.
- Andaria, M., & Hadiwinarto. (2020). Development of a two-tiier multiple choice question assessment instrument to measure students science process skills on acid-base material. *ISEJ: Indonesian Science Education Journal*, 1 (3), 257-268. <https://siducat.org/index.php/isej/article/view/141>.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi/ Journal of Mathematics, Statistics, & Computing*, 9(2), 95-113. <https://doi.org/10.20956/jmsk.v9i2.3402>.
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Jakarta: Bumi Aksara.
- Bloom, B. S. (1956). *Taxonomy of Educational Objectives, Handbook: The Cognitive Domain*. New York: David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice* , 8 (3), 293-307. <https://doi.org/10.1039/B7RP90006F>.

- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt: New York.
- DeMars, C. E. (2010). *Item response theory*. New York: Oxford University Press.
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Hong Kong: Assessment & Support Team. <https://wlts.edb.hkedcity.net/en/home/AandLI2.html>.
- Gierl, M. J., Bulut, . O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research* , 87 (6), 1082–1116. <https://doi.org/10.3102/0034654317726529>.
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. England: Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice* , 12 (3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543>.
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage Publications.
- Jamhari, M. (2021). Developing the two-tier multiple choice tests in enhancing students' higher-order thinking skills on human digestive system. *Eduproxima: Jurnal Ilmiah Pendidikan IPA* , 3 (1), 50-64. <https://doi.org/10.29100/eduproxima.v3i1.1853>.
- Kartono. (2008). Equating the combined dichotomous and polytomuos item test model in an achievement test. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>.
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus. *5th ICRIEMS Proceedings* (pp. 479-485). Yogyakarta, Indonesia: Faculty of Mathematics and Natural Sciences. <http://seminar.uny.ac.id/icriems/proceeding2018>.
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://doi.org/10.31327/jme.v6i2.1607>.
- Linacre, J. M. (2012). *Winstep : Rasch-model computer programs*. Chicago: Winsteps.Com. <https://www.winsteps.com/winman/copyright.htm>.

- Malhotra, N. K. (2006). *Riset Pemasaran* [Marketing Research]. Jakarta: Eirlangga.
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *Current research in Pandemic Covid-19 Era in Indonesia - 2nd ICOSETH 2020* (pp. 44-55). Surakarta, Indonesia: University of Sebelas Maret. <https://doi.org/10.20961/ijsascs.v4i1.49457>.
- Plomp, J. (2013). *Educational design research: An introduction*. Netherlands: Netherlands Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>.
- Prasetya, A., Rosidin, U., & Herlina, K. (2019). Development of instrument assessment for learning the polytomous response models to train higher order thinking. *Young Scholar Symposium on Transdisciplinary in Education and Environment (YSSTEE)* (pp. 1-11). Bandar Lampung: UIN Raden Intan. <https://doi.org/10.1088/1742-6596/1155/1/012032>.
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://doi.org/10.15294/jere.v9i2.46133>.
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Yogyakarta, Indonesia: Nuha Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>.
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Yogyakarta: Media Akaademi.
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst : Jurnal Pendidikan Fisika dan Terapannya/ Karst : Journal of Physics Education and Its Application*, 4 (1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>.
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam questions for Islamic religious education and character at the elementary school level]. *An-Nahdhah/ An-Nahdhah* , 11 (2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article/view/40>.
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Classical tes theory vs response theory? *Didaktika: Jurnal Kependidikan/Didactics: Journal of Education* , 13 (1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>.
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research* , 10 (3),

1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan* [Rasch Modeling Applications in Educational Assessment]. Yogyakarta: Trim Komunikata.

Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic tests in learning]. Yogyakarta: Graha Ilmu.

Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research* , 9 (4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>.

Treagust, D. (1988). Development and use of diagnostic tests to evaluate student's misconceptions in science. *International Journal of Science Education* , 10 (2), 159-169.

Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep* [Analysis of research data with rash and winstep models]. Bengkulu: Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122-1?inline=1>.

Widiyatmoko, A., & Shimizu, K. (2018). The Development of two-tier multiple choice test to assess students' conceptual understanding about light and optical instruments. *Jurnal Pendidikan IPA Indonesia* , 7 (4), 491-501. <https://doi.org/10.15294/jpii.v7i4.16591>.

Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. Bandung: UPI Press.

Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology* , 3 (1), 60-80. <https://doi.org/10.14742/ajet.2154>.



KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS LAMPUNG
UPT BAHASA

Jalan Prof. Dr. Soemantri Brojonegoro No. 1 Bandar Lampung 35145
Telepon : (0721) 770844, Whatsapp : 0811 724 5544, email : uptbahasa@kpa.unila.ac.id
Website : uptbahasa.unila.ac.id



CERTIFICATE OF PROOFREADING

Number: 047/UN26.33/TU.00.08/2022

The undersigned below,

Name : Dr. Muhammad Sukirlan, M.A.



NIP : 196412121990031003

Position : Head of Language Center - University of Lampung

states that the article entitled : **"DEVELOPMENT OF ASSESSMENT FOR LEARNING INSTRUMENT USING POLYTOMOUS RESPONSE IN MATHEMATICS IN VOCATIONAL SCHOOLS"** written by Sugeng Sutlarso, Undang Rosidin, Aan Sulistiawan has been edited or proofread in terms of linguistic aspects by Language Center University of Lampung.



Bandar Lampung, February 18th, 2022
Head,


Dr. Muhammad Sukirlan, M.A.
NIP 196412121990031003 

check. We will inform you when we get the result from our reviewers.

If the reviewers confirm your revised paper, we will send the acceptance letter to you.

Thank you for your patience.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

[Quoted text hidden]



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

2nd round corrections request for the manuscript ID# 21112502244011

6 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>
Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Mon, Feb 28, 2022 at 7:59 PM

Dear Dr. Sutiarso,

Please see the attached file as the second round corrections.

Please remove the old highlights and re-highlight for new edited parts. We need a new correction report.

We are looking forward to getting your second revised paper until **March 14, 2022**.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com



On 22-Feb-22 4:48 PM, SUGENG SUTJARSO wrote:

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.

I have revised the article according to the reviewer's suggestion. Here I attach (1) a revised article, (2) a correction report, and (3) a proofreading certificate from my university's language center.

Best regards,

Sugeng Sutiarso
Lampung University

2 attachments **2nd round__EU-JER ID# 21112502244011_R2612.dot**
170K **2nd round__EU-JER ID# 21112502244011_R2613.docx**
1573K

SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Sun, Mar 13, 2022 at 7:51 PM

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.I have revised the second round of corrections according to the reviewer's suggestion.
Here I attach a correction report and revised article.

Thank you.

Best regards,

Sugeng Sutiarso
Lampung University

[Quoted text hidden]



European Journal of Educational Research

ISSN: 2165-8714

<http://www.eu-jer.com/>

Review Form

Manuscript ID:	EU-JER_ID#_21112502244011	Date: 28 February 2022
Manuscript Title:	Development of assessment for learning instrument using polytomous response In mathematics in vocational schools	

ABOUT MANUSCRIPT (Mark with "X" one of the options)	Accept	Weak	Refuse	Not Available
Language is clear and correct		X		
Literature is well written		X		
References are cited as directed by APA	X			
The research topic is significant to the field		X		
The article is complete, well organized and clearly written		X		
Research design and method is appropriate	X			
Analyses are appropriate to the research question	X			
Results are clearly presented	X			
A reasonable discussion of the results is presented		X		
Conclusions are clearly stated		X		
Recommendations are clearly stated		X		

GENERAL REMARKS AND RECOMMENDATIONS TO THE AUTHOR

Title

The title should be "The development of an assessment instrument using Polytomous response in mathematics"

Abstract

Please use past tense in the abstract. For example, write showed instead of show. Use "Questionnaire was to identify" instead of "Questionnaire is to identify".

Introduction

The revisions to my previous comments are not satisfactory. The introduction was shortened but its current version does not still explain what the problem is. From the authors' writing, we do not have adequate information about the strengths and weaknesses of the existing instruments. The explanations are superficial.

My previous questions remain valid. What do we know about the existing instruments? What do we need to know about the existing instruments?

Method

Use the past tense "The instrument is validated by two people..."

Discussion

I could not understand "The similarity of Sarea's research is safe..."

The word "the product" is not acceptable in an educational paper. - "Although the products are different ...".



European Journal of Educational Research

ISSN: 2165-8714

<http://www.eu-jer.com/>

Conclusion

The new knowledge is not explained in the literature?

Rewrite – “the condition that must exist so that an instrument with a response polytomous can be used to determine students' abilities is the suitability of the results of the analysis between classical and modern theories”

Language

The use of the language is still problematic.

Test

The developed instrument should be given in an appendix.

THE DECISION (Mark with "X" one of the options)

Accepted: Correction not required	
Accepted: Minor correction required	
Conditionally Accepted: Major Correction Required (Need second review after corrections)	X
Refused	

Reviewer Code: R2612 (The name of referee is hidden because of blind review)

DEVELOPMENT OF ASSESSMENT FOR LEARNING INSTRUMENT USING POLYTOMOUS RESPONSE IN MATHEMATICS IN VOCATIONAL SCHOOLS

Abstract: This research is development research that aims to produce a good instrument of assessment in mathematics using polytomous response according to classical and modern theories. Plomp's model as a research design has five stages: preliminary investigation, design, realization or construction, trial, revision, and implementation (testing). The study is conducted in three vocational schools in Lampung, Indonesia. The study involves 413 students consisting of 191 male students and 222 female students. The data are collected through questionnaires and tests. Questionnaire is to identify instruments commonly used by teachers so far and to validate instruments by experts. The test uses multiple-choice tests with open reasons as many as 40 items. The data are analyzed in two ways, namely analysis with classical and modern theories. The results show that the instrument of assessment in mathematics with polytomous response has good categories according to classical and modern theory, although item discrimination according to classical theory needs to be revised, and the instrument of assessment in mathematics with the polytomous response (multiple-choice tests with open reason) can provide information on the actual competency of students. It is evidenced by the suitability of the results of the analysis of classical and modern theories.

Keywords: assessment for learning, classical and modern theory, multiple choice tests with open reason, polytomous response, vocational school

Introduction

Assessment is an important activity and needs to be done by teachers in schools. The conventional paradigm often interprets the assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). If referring to the current paradigm, assessment in schools is divided into three parts, namely assessment as learning,

assessment for learning, dan assessment of learning (Wulan, 2018). Assessment as learning has almost the same function as assessment for learning but assessment as learning involves students in assessment, such as assessing themselves or colleagues. Assessment of learning is an assessment carried out after all learning ends and aims to assess student achievement. Assessment for learning is an assessment carried out during the learning process and aims to improve learning. It can play a role in preventing students from experiencing further learning failure because of its position between the other two assessments (Earl, 2013) as shown in the following assessment pyramid (Figure 1).



Figur 1. Assesement pyramid

Assessment activities can be applied with tests. A test is a tool or procedure used to find out or measure students' abilities about something with certain rules (Arikunto, 2012). According to the type, the test consists of two types, namely multiple choice and essay. Each form has advantages or disadvantages to each other. The advantage of multiple-choice over essays is that multiple-choice can measure multiple cognitive levels, scoring more objectively, and saving time. Meanwhile, essay tests can only describe students' actual abilities, but scoring tends to be subjective (Rosidin, 2017). The results of the preliminary research survey found that so far teachers have used multiple choice tests to determine students' abilities. During the test, students tend to guess on difficult items, sometimes students guess right and sometimes wrong. Of course, it is difficult to distinguish between students who answered correctly and guessed. Therefore, it is necessary to develop multiple choice tests that prevent students from guessing the answers.

Specifically for choice tests, in the last four decades, evaluation experts have developed multiple-choice tests, namely multiple-choice tests with open and closed reasons(Suwarto, 2012). Multiple choice test with open reasons is a multiple-choice test that has alternative answers and questions are asked for the choice of choice. Meanwhile, multiple-choice tests with closed reasons are multiple-choice tests that have alternative answers, and students are asked to choose the reasons provided. The advantage of the choice test, compared to the free choice is the answer from the answer chosen, compared to the multiple tests which are not free to give reasons. Generally, multiple-choice tests on assessment for learning do not provide reasons (open or ask closed). The multiple-choice test has only true or false answers. If the answer is correct, then it gets a score of 1. Otherwise, if the answer is wrong, it gets a score of 0, and this scoring is called dichotomous. The choice test provides a choice of opportunities for students to get a score even though the answer is wrong and this scoring is called polytomous(Kartono, 2008).

The first time, the polytomous response test was in the 80s, and the test was known as the two-tier multiple-choice test(Treagust, 1988). The test is in the form of multiple-choice with closed reasons which aim to diagnose errors in the concepts of biology, physics, and chemistry. Multiple choice test with closed reasons has two levels, the first level is a multiple-choice test and the second level is a multiple-choice test with a choice of reasons from the first level questions(Chandrasegaran, et al., 2007). Furthermore, studies on the development of multiple-choice tests with closed reasons have been carried out by many researchers on several concepts, such as physics on light and optical instruments (Widiyatmoko & Shimizu, 2018), mathematics on calculus(Khiyarunnisa & Retnawati, 2018), chemistry on acids and base (Andaria & Hadiwinarto, 2020), biology on the respiratory system(Myanda, et al., 2020), mathematics on reasoning problems(Ambarwati, et al., 2020), biology on the human digestive system(Jamhari, 2021), mathematics on

mathematical connections(Lestari, et al., 2021). Meanwhile, the development of multiple-choice tests with open reasons is still limited, namely mathematics in calculus(Yang, et al., 2017), and outside mathematics, namely physics in Higher Order Thinking Skills(Prasetya, et al., 2019).

There are two approaches to analyzing test items, namely classical and modern theory. Classical theory has weaknesses. It cannot separate the characteristics of respondents and test items. It means that the test taker's ability is only determined by the test. The characteristics of the test items will change when the examinees change, and the characteristics of the examinees will change when the items change. Thus, classical theory is considered less able to provide information on the actual abilities of students because the results of the assessment are highly dependent on the respondents. Modern theory, with item response theory, is a solution to overcome the weakness of classical theory because item response theory has the concept of releasing the relationship between respondents and test items(Saepuzaman, et al., 2021).It means a test that has a good category not only according to classical theory but must be supported by test analysis according to modern theory.

Classical theory is a measurement theory for assessing tests based on the assumption of measurement errors between actual results and observations, or correlation measurement errors based on test-takers. From the assumption, a formula for calculating the level of difficulty and item discrimination was developed(Hambleton & Jones, 1993). The modern theory is a theory of measurement to assess by comparing the average performance against the appearance of evidence of group ability, or the ability of test-takers predicted from their abilities; or better known as Item Response Theory/IRT(Hambleton & Linden, 1982). Classical test theory is used because it is easy to apply but has limitations in measuring the level of difficulty and differentiating items because the calculation of the two indicators is

based directly on the total score of the test takers. While modern theory with item response theory frees the dependence between test items and respondents (parameter invariance concept), responses to one test item do not affect other test items (local independence concept), and test items only measure one dimension or unidimensional concept (Anisa, 2013).

Related to the polytomous response test, a measurement innovation is needed that can guarantee a good test according to classical theory and modern theory. This research is development research that aims to produce a good instrument of assessment in mathematics using polytomous response according to classical and modern theories. If the results of this research are obtained a proper test according to classical and modern theory, it can make it easier for teachers to develop appropriate tests on other materials, and students can find out their true abilities. Finally, for practitioners or other researchers, it can be a reference in further research. The formulation of the problem proposed is (1) does the instrument in the assessment in mathematics developed with the polytomous response have a good test category according to classical and modern theory measurements?, and (2) does the assessment for learning an instrument with the polytomous response provide information on the actual competence of students?

Research Method

Research Design

This research is a research and development that refers to Plomp's (2013) model with the research procedure consisting of five stages, namely preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

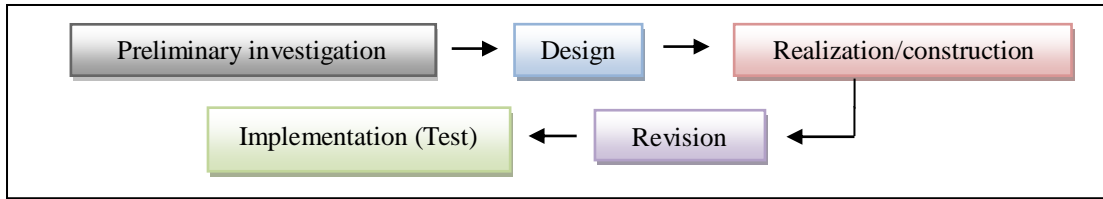


Figure 2. Stages of research design

The preliminary investigation stage is to identify the assessment instruments used by teachers so far. The design stage is to make a polytomous response test grid according to the basic competencies of mathematical concepts and to make an expert assessment questionnaire sheet. The realization or construction stage is developing the items on the polytomous response test, and also the expert validation process for the items on the developed polytomous response test. The revision stage is the improvement of items on the polytomous response test based on expert advice. The implementation (test) stage is to try out the polytomous response test to students and analyze the results of the test.

Research Subject

The subjects of the study are students of a vocational school in the province of Lampung, Indonesia. The research sample is determined using a non-probability sampling technique in the form of accidental sampling. It means taking the subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools are three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects are 413 students at grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 from the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
SMK Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collecting Technique

Data are collected using a questionnaire and test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed and to determine content validity (Suhaini, et al., 2021). The instrument is validated by two people who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert are the suitability of the items with the indicators, language, and alternative answers to the questions. The score in validating the instrument follows the criteria as in Table 2 below.

Table 2. Content validity score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining content validity, the instrument is tested on students. Then, it is continued by determining the validity of the construct and its reliability, with the aim that the instrument can be further analyzed.

The instrument used is a multiple choice test with a polytomus response with open reasoning,

which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contains five answer choices along with the reasons. Student answers score refers to the polytomus score in the Partial Credit Model, where answer choices and reasons are related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Student answers score

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The research data obtained are analyzed in two stages, namely (1) questionnaire data analysis (qualitative analysis), and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis.

1. Questionnaire data analysis (qualitative analysis)

The questionnaire data analyzed included two parts, namely identification data on the instruments used by the teacher and expert validation data. The identification data on the instrument are analyzed descriptively, and the expert validation data are analyzed for trends or expert agreement using the Gregory formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range 0 - 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the value of the validity of an item.

2. Test data analysis (empirical analysis)

After the content validity test, it is followed by construct validity and reliability tests. The construct validation test uses exploratory factor analysis. The instrument is considered to have a good construct if the explained Kaiser-Meyer-Olkin (KMO) value is more than 0.5 (Retnawati, 2014). Reliability test using Cronbach's Alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further tests can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is because they are both preliminary analyzes of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program is used for classical theory, and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program is used because it has several advantages (Untary, et al., 2020), namely, it can analyze polytomous data and can analyze the maximum likelihood model using a 1-parameter logistic model.

2.1 Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answer the questions correctly or incorrectly. The difficulty of a good (medium) item is if the index is in the range of 0.3 to 0.7. If the index is below 0.3 then the item is difficult, and vice versa if the index is above 0.7 then the item is easy.
- b. Item discrimination is the item's ability to distinguish high-ability students from

stupid, low-ability students. Good item discrimination if it has an index above 0.3; and if the item discrimination index is below 0.3 then the item needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

a. The item difficulty level is the level of the student's latent trait towards the item.

The difficulty of the items identifies the ability of about 50% of respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.

b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column.

According to modern theory, before analyzing the level of difficulty and item discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton, et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigen values of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals on the unidimensional solution (DeMars, 2010). In this study, the dimensional test uses the Eigen value analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer on each item. If the unidimensional assumption has been met, the local independence assumption has also been met (DeMars, 2010). Model fit test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring Outfit Mean Square (MNSQ) and Pt-Measure. If the Outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and Standard Error Measurement (SEM) are analyzed which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

1. Analysis of Questionnaire Data

Based on the results of the questionnaire, it is found that so far the teacher has never used the polytomous response instrument with a multiple-choice test with open reasons. As many as 80% of teachers use essay tests and 20% of teachers use multiple-choice, with each instrument consisting of 2-5 items. In addition, about 10% of teachers use this assessment as a learning improvement, such as improving lesson plans and teaching models/methods. The results of the questionnaire stated that 90% of teachers who do not use assessment as an improvement in learning are caused by several aspects, such as teachers are not understanding assessment (20%), teachers are not knowing how to analyze assessments (50%), and teachers are not knowing how to develop good assessment questions (30%). The following is a summary of the questionnaire data from the identification of assessment for learning instruments.

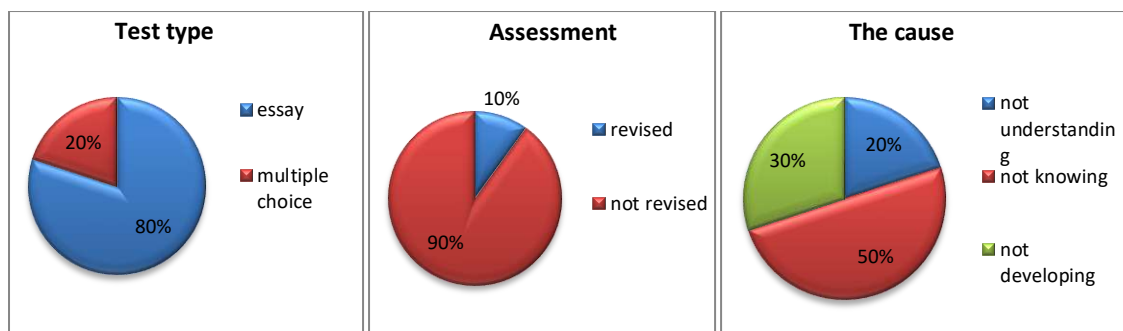


Figure 3. Description of teacher condition in assessment for learning

Content Validity

The results of the two expert assessments show that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement with the Gregory Index formula is obtained as shown in Table 4 below.

Table 4. Index Gregory items

		Rater 1	
		weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provide some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arrange them in order.

2. Analysis of Test Data

Construct Validity

After testing the instrument, it is followed by a construct validity test. The results of the test

with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory factor analysis

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained the Cronbach's Alpha coefficient value of 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to the classical and modern methods.

Table 6. Item reliability

Cronbach's Alpha	N of Items
0.892	40

2.1 Analysis of Test Data with Classical Theory

Analysis of test data with classical theory does not require testing assumptions, but the analysis of the level of difficulty and item discrimination can be directly calculated. The results of the analysis of the level of difficulty and item discrimination are obtained as shown in Table 7 below.

Table 7. Level of difficulty and item discrimination

Item	Level of difficulty	Category	Discrimination	Category	Item	Level of difficulty	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised
4	0.540	good	0.304	Good	24	0.438	good	-0.071	revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	revised

13	0.400	good	-0.321	Revised	33	0.441	good	0,036	revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	revised

Based on Table 7, it is found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are included in the good category. Meanwhile, only two items on discrimination have good categories, and the remaining items need to be revised. The results indicate that all items are good based on the level of difficulty, but almost all items need to be revised for item discrimination.

2.2 Analysis of Test Data with Modern Theory

Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, then calculating the eigenvalues. The eigenvalue is then used to calculate the percentage of explained variance, as well as describe the scree plot (Retnawati, 2014). The output of factor analysis is the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. KMO test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

The unidimensional test is based on the cumulative percentage of eigenvalues and scree plots. If the cumulative percentage of eigenvalues in the first factor is more than 20%, then the unidimensional assumption is accepted (Retnawati, 2014). In Table 9, it can be seen that the

cumulative percentage of the eigenvalues in the first factor is 20.220%. The cumulative percentage of the eigenvalues has exceeded 20%,so the instrument in the study is proven to only measure one factor or dimension.

Table 9. Total variance explained

Component	Initial Eigenvalues		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot which is based on the number of factors marked by the steepness of the graph with the acquisition of eigenvalues.

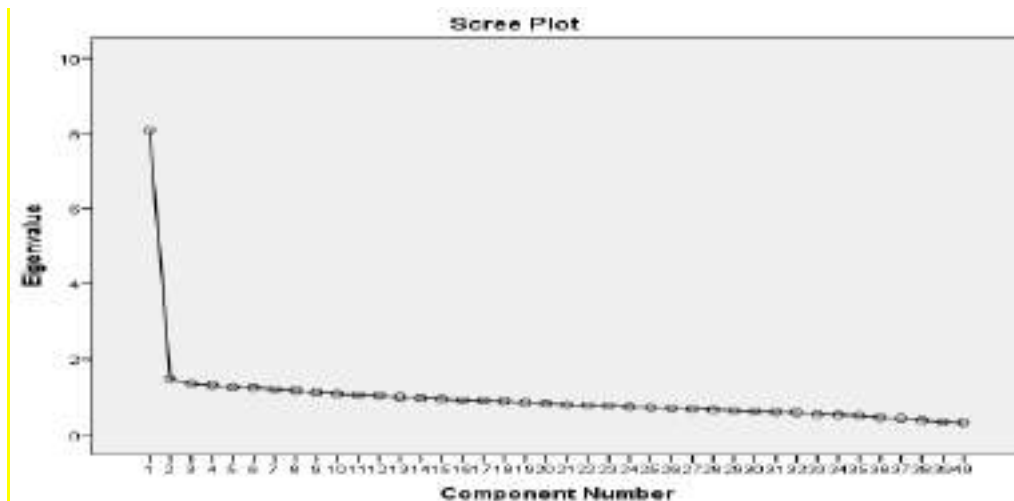


Figure 4. Scree plot unidimensi

Based on the scree plot, it is known that the eigenvalues immediately slope on the second factor. It shows that there is only one dominant factor in the developed instrument. The results prove that the test kit meets the unidimensional assumption or in other words only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be accepted if the respondent's answer to one item does not affect the respondent's answer to another item. Thus, the score of one item

should not be determined or dependent on the scores of other items. It confirms that the assumption is automatically proven after being proven by the unidimensionality of the respondent's data on a test (Retnawati, 2014).

Table 10. Covariance matrix

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Data in Table 10 shows the results of the variance-covariance values between groups of students' abilities. It can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. It can be concluded that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test is analyzed using the Winsteps program. The item requirements are called fit to the model if the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Corr is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items match the model or fit (Table 11).

Table 11. Item fit on model

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73
 Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.4	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	-.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	-.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Difficulty Level

The item difficulty level is analyzed using the Winsteps program, and the results obtained can be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of -2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely the difficulty level is in the range of -2 and 2.

Table 12. Item difficulty level

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73
 Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ ZSTD	OUTFIT MNSQ ZSTD	PT-MEASURE CORR.	EXP.	EXACT MATCH OBSN	EXPN	Item
1	1134	413	-.70	.07	1.16 2.5	1.16 2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97 -4	.98 -3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93 -1.0	.94 -1.0	.67	.43	49.4	49.3	Q3
4	1064	413	-.38	.07	.82 -3.0	.82 -3.0	.46	.43	50.1	48.6	Q4
5	1021	413	-.19	.07	.84 -2.6	.84 -2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16 2.4	1.16 2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03 -3	1.03 -3	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16 2.4	1.16 2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97 -5	.97 -5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93 -1.0	.93 -1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97 -4	.98 -4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90 -1.7	.90 -1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07 1.1	1.07 1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01 2.1	1.01 3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10 1.5	1.10 1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08 1.2	1.08 1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96 -7	.96 -6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12 1.8	1.12 1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04 -7	1.05 -8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97 -4	.97 -4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01 -3	1.02 -3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08 1.2	1.08 1.2	.56	.44	45.6	50.0	Q22
23	1037	413	-.26	.07	1.05 8	1.05 8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06 1.0	1.07 1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00 -0	1.00 -0	.41	.44	47.9	51.1	Q25
26	943	413	-.16	.07	.97 -4	.97 -4	.31	.44	50.8	50.9	Q26
27	936	413	-.20	.07	1.02 -3	1.02 -4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89 -1.8	.88 -1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97 -5	.97 -5	.33	.44	49.2	51.0	Q29
30	949	413	-.14	.07	1.06 -9	1.06 -9	.30	.44	45.5	50.9	Q30
31	955	413	-.11	.07	1.07 1.1	1.07 1.1	.25	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12 1.8	1.12 1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04 -6	1.04 -6	.41	.44	48.2	51.0	Q33
34	959	413	-.09	.07	1.06 1.0	1.07 1.1	.33	.44	46.5	51.0	Q34
35	955	413	-.11	.07	1.04 -7	1.04 -7	.40	.44	47.2	51.0	Q35
36	961	413	.05	.07	1.04 -8	1.05 -8	.34	.44	47.7	51.0	Q36
37	955	413	-.11	.07	.96 -6	.96 -6	.40	.44	51.1	51.0	Q37
38	921	413	-.27	.07	.89 -1.7	.89 -1.8	.43	.44	52.8	50.6	Q38
39	825	413	-.72	.07	.85 -6.3	.86 -6.2	.59	.43	58.1	48.6	Q39
40	801	413	-.84	.07	.74 -4.5	.74 -4.6	.54	.41	55.4	48.2	Q40
MEAN	979.5	413.0	-.00	.07	1.00 -1.1	1.00 -1.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11 1.8	.11 1.8			4.3	.8	

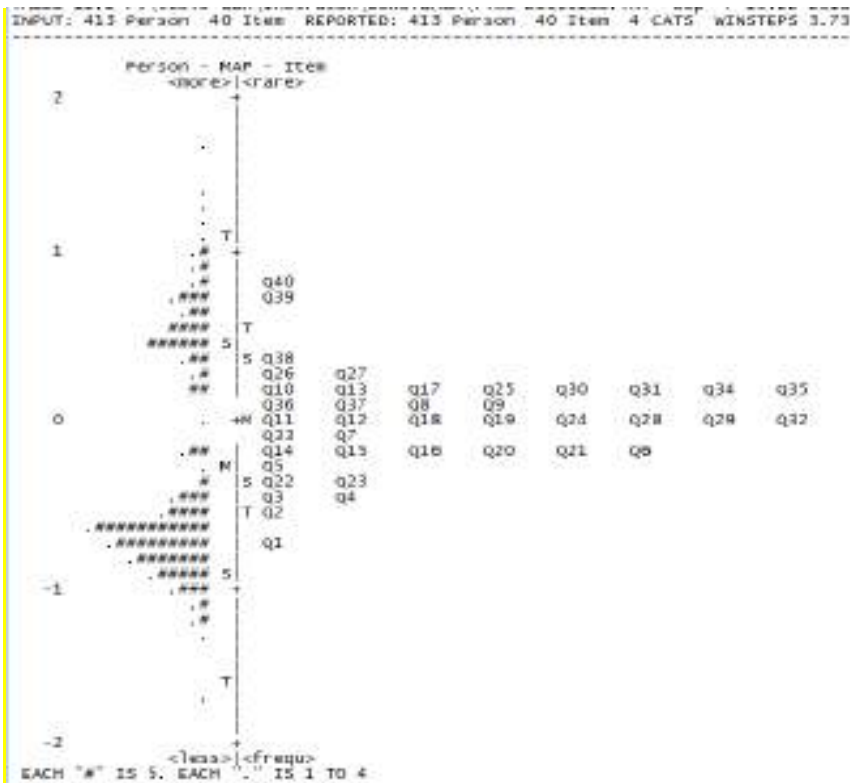


Figure 5. Item difficult map

Item Discrimination

The item discrimination analysis used the Winsteps program and the results are presented in Table 13 (column PT-Measure Corr). Item discrimination is in the range of 0.23 to 0.74; with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai, et al., 2005).

Table 12. Item discrimination

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73
 Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Comparative Analysis between Classical and Modern Theory

The results of the analysis of classical and modern theory obtained the index of difficulty level and item discrimination as follows.

Table 14. Analysis classical and modern theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Category	Percentage	Many Items with Good Category	Percentage
difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 14, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in good category than using classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Table 13), it can be seen that there is a match between the categories of item discrimination. It means that if the item discrimination is not good with the classical theory, then the item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability is measured by using a test that is expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error or Standard Error Measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

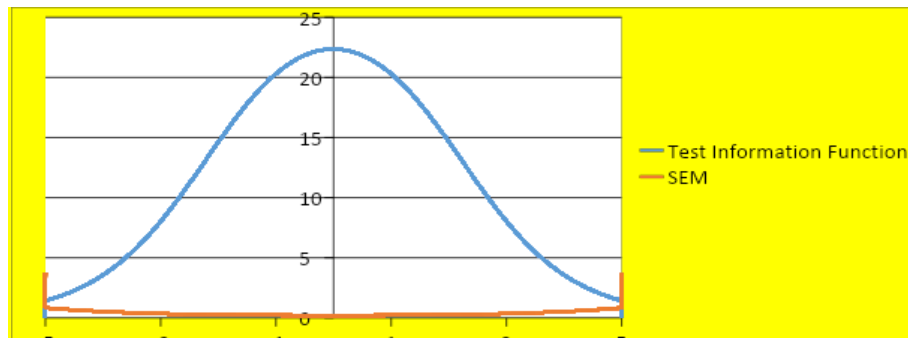


Figure 5. Graph of information function and measurement error

Figure 5 shows that the instrument provides a maximum of 22.36 information and has the smallest measurement error of 0.21 if it is given to students with moderate ability, which is 0.2. The lower limit and upper limit of the interval is the ability score where the graph of the information function and the SEM graph intersect in that interval. The graph indicates that the greater the value of the information function, the smaller the measurement error (SEM). Item information function states the strength or contribution of test items in revealing the latent trait as measured by the test. With the item information function, it is known which items match the model, thus helping in the selection of test items (Retnawati, 2014). In conclusion, the characteristics of the test kit are suitable for students with moderate abilities.

Based on the test results, in addition to knowing the quality of the developed test instruments, it can also be seen the ability of students to work on the questions given. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of 6 student answers were selected as samples with different abilities (high, medium, and low).

Item 1: the cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform

algebraic operations on general forms. arithmetic sequence.

Question 1:	Pattern 1:	Pattern 2:
<p>Diketahui barisan bilangan 3, 8, 13, 18,.... Rumus suku ke-n barisan itu adalah....</p> <p>a. $U_n = 5n - 3$ b. $U_n = 5n - 2$ c. $U_n = 2n + 1$ d. $U_n = 4n - 1$ e. $U_n = 3n + 2$</p> <p>Alasan:</p>	<p>1. diketahui : $a = 3$ $b = 5$ ditanya $U_n = ?$ jawab : $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	<p>$U_n = a + (n-1)b$ $b = u_2 - u_1 = 8 - 3 = 5$ $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>

Figure 6. Student answers in item 1

Item 2: the cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand and determine the number of terms in a sequence by using the general formula for an arithmetic sequence or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

Question2:	Pattern 1	Pattern 2
<p>Diketahui barisan aritmetika 2, 5, 8, 11, ...68. Banyaknya suku barisan tersebut adalah....</p> <p>a. 12 b. 13 c. 22 d. 23 e. 24</p> <p>Alasan:</p>	<p>2, 5, 8, 11, ... 68 02, 05, 08, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 59, 62, 65, 68 Jadi banyak suku tersebut adalah 23</p>	<p>$68 = a + (n-1)b$ $68 = 2 + (n-1)3$ $68 - 2 = (n-1)3$ $66 = (n-1)3$ $21,8 = 1,5n$ $n = 14,5$ (.)</p>

Figure 7. Student answers in item 2

Item 3: the cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and determined the difference or difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the

terms of the known terms and inserting several terms.

Question 3:	Pattern 1	Pattern 2
<p>Pada suatu barisan aritmetika diketahui suku Ke-5 adalah 31, dan suku ke-8 adalah 46. Nilai beda barisan tersebut adalah</p> <p>a. 5 b. 6 c. 7 d. 8 e. 11</p> <p>Alasan:</p>		

Figure 8. Student answers in item 3

Item 4: the cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and can determine the n th term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula, but by writing the terms of the known terms and inserting several terms and continuing until the n th term.

Question 4:	Pattern 1	Pattern 2
<p>Jika barisan aritmetika berturut-turut adalah 110 dan 150. Suku ke-30 barisan aritmetika tersebut adalah</p> <p>a. 308 b. 318 c. 326 d. 344 e. 354</p> <p>Alasan:</p>		

Figure 9. Student answers in item 4

Item 5: the cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but by writing the terms from known terms and inserts several terms and then defines them.

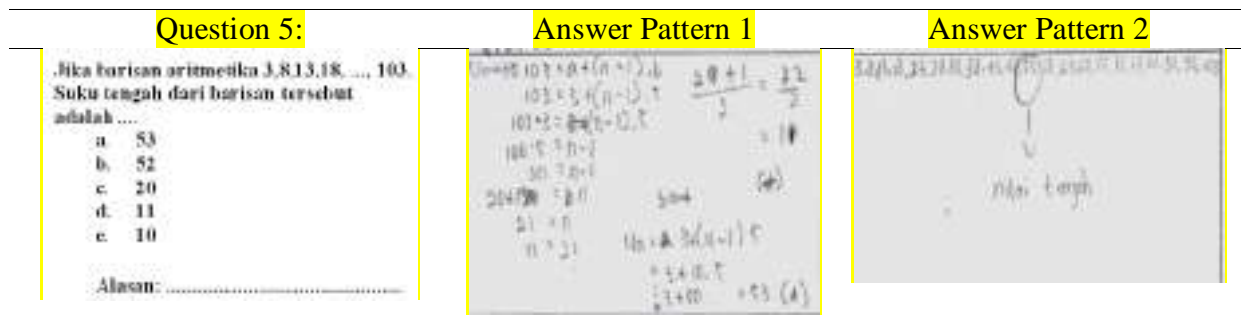


Figure 10. Student answers in item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is a multiple-choice test with open reasons and all parameters have been accepted. This instrument is a combination of multiple-choice test and essay. Multiple-choice tests are easier to check students' answers but students' mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find out deeper mathematical thinking processes but it takes a long time to check the answers.

Assessment of learning instruments that have been carried out with item analysis and the results of student analysis is one of the important sources of composite scores to be reported. In the final report, the test taker's ability score should be changed to a score of 0 - 10 from 0 - 100, according to the needs of the school. The transformation uses a linear transformation by dividing the score by the ideal score and then the result is multiplied by 10 to get a value in the range 0 – 10 or multiplied by 100 to get a score of 0 – 100. In the range 0 – 10, the score obtained by students taking the test the highest mathematics learning was 8.56 and the lowest was 4.31. In the range 0 – 100, the scores obtained by students with the highest mathematics is 85.625 and the lowest is 43.125.

The results of the assessment of students' mathematical abilities are presented in the form of very low to very high predicates. The results of the test analysis show that most students have low and very low abilities, namely 62% (253 students). Meanwhile, students who have high

and very high abilities are 38% (160 students). The results of another analysis find that students who have high abilities tend to work according to the concepts that have been given by the teacher but do not follow the completion steps, students who have moderate abilities can solve problems according to the concepts that have been given by the teacher and the steps, and there are students who have abilities but are not able to use the concepts given by the teacher and are not even able to give clear reasons..

Another result of this study is that teachers agree to provide learning assessments with multiple choice questions with open-ended reasons because the instrument is easier for teachers to find out students' difficulties in certain materials. In this way, teachers can also provide remedial or other assistance to students who have learning difficulties. It means that the polytomous response instrument can be used as a way to determine which students need remedial or not. In general, previous research stateshow to determine students who need remedial only one test, namely multiple-choice tests(Gierl, et al., 2017)or essays(Putri, et al., 2020).

Discussion

This research is development research to produce an instrument using polytomous response. The instrument is a multiple-choice test with open reasons. This instrument is analyzed by classical and modern theory. There are differences in the results of the analysis between classical and modern theories, namely item discrimination. Classical theory analysis obtains 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtains 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as not good but the results of the modern analysis are categorized as good, and vice versa(Retnawati, 2014). It means, if you

find items that are not in a good category with classical theory, do not be in a hurry to revise or replace them before the analysis of modern theory analysis.

Research on assessment for learning with polytomous responses with multiple-choice questions (having open reasons) is still limited. When compared to previous research, there is only one study on assessment for learning with polytomous responses to open-ended multiple-choice questions (Yang, et al., 2017). However, Yang study has several fundamental differences, namely the research objectives and data analysis. The research aims to diagnose student errors in university on the concept of calculus, not to produce a good assessment instrument. The data analysis uses parametric statistics (covariance), not using item analysis (classical and modern). Since the objectives and data analysis are different, the results of the study cannot be compared with the results of this study. However, this research has provided a reference for researchers in making reasoned multiple-choice questions, such as the suitability of items with indicators, language, and alternative answers to questions.

Other studies are similar to assessment for learning with polytomous responses on multiple-choice questions with reasons (Sarea, 2018). The similarity of Sarea's research is safe, it lies in the research objectives and the analysis used (classical and modern). However, the difference is the researchers do not develop their questions and the questions are in the form of closed multiple-choice questions. The results of Sarea's research states that the comparison of the results of the classical and modern methods of item analysis is different. The difference is that the level of difficulty and item discrimination in the classical method is more categorized as good than the modern method. In other words, the modern way of stating the level of difficulty and item discrimination is categorized as good even though the analysis method states that the items are categorized as not good. Likewise with Saepuzaman's research, namely the product developed in the form of multiple-choice questions with closed

reasons. Although the products are different, the results of the research can contribute to this research. The contribution of both is that there is a belief in the items that are declared not good by classical analysis, which turn out to be good with modern analysis. The results of previous studies have provided support to the results of the research that the polytomous response instrument in the form of multiple-choice questions with open reasons can be used as an alternative assessment for learning, as well as other assessments (assessment as learning and assessment of learning) for all vocational schools in Lampung, Indonesia and even outside Indonesia.

Conclusion

Based on the results of the research and discussion, conclusions are obtained, namely (1) the instrument with a polytomous response has been accepted according to classical and modern theory, although item discrimination in the classical theory needs to be revised; **It means if you find items that are not in a good category with classical theory, do not be in a hurry to revise or replace them before the analysis of modern theory analysis**, and (2) an instrument with a polytomous response (multiple-choice test with open reasons) can provide information on actual student competencies. It is evidenced by the suitability of the results of classical and modern analysis. **So, the condition that must exist so that an instrument with a response polytomus can be used to determine students' abilities is the suitability of the results of the analysis between classical and modern theories.**

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, teachers should familiarize students with giving a test in the form of a polytomous response before giving the test. For schools, principals or other leaders should encourage other teachers to take advantage of this test, and develop other assessment

instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomus (pretest). This is important, so that students' prior knowledge can be known so that learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not been the researchers' expectations, for example representing schools with high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic material (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Dordrech: Springer-Kluwer Academic Publishers.
- Ambarwati, R., Sunardi, Yudianto, E., Murtikusuma, R. P., & Safrida, L. N. (2020). Developing mathematical reasoning problems type two-tier multiple choice for junior high school students based one thnomathematics of jember fashion carnival. *ICOLSSTEM*. Jember: IOP Publishing. <https://doi.org/10.1088/1742-6596/1563/1/012036>.
- Andaria, M., & Hadiwinarto. (2020). Development of a two-tiier multiple choice question assessment instrument to measure students science process skills on acid-base material. *ISEJ: Indonesian Science Education Journal*, 1 (3), 257-268. <https://siducat.org/index.php/isej/article/view/141>.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi/Journal of Mathematics, Statistics, & Computing*, 9(2),95-113. <https://doi.org/10.20956/jmsk.v9i2.3402>.

- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Jakarta: Bumi Aksara.
- Bloom, B. S. (1956). *Taxonomy of Educational Objectives, Handbook: The Cognitive Domain*. New York: David McKay. Title in sentence case
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8 (3), 293-307. <https://doi.org/10.1039/B7RP90006F>.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt: New York.
- DeMars, C. E. (2010). *Item response theory*. New York: Oxford University Press.
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Hong Kong: Assessment & Support Team. <https://wlts.edb.hkedcity.net/en/home/AandLI2.html>.
- Gierl, M. J., Bulut, J. O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87 (6), 1082–1116. <https://doi.org/10.3102/0034654317726529>. Delete yellow highlighted comma and dot
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. England: Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12 (3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543>.
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage Publications.
- Jamhari, M. (2021). Developing the two-tier multiple choice tests in enhancing students' higher-order thinking skills on human digestive system. *Eduproxima: Jurnal Ilmiah Pendidikan IPA*, 3 (1), 50-64. <https://doi.org/10.29100/eduproxima.v3i1.1853>.
- Kartono. (2008). Equating the combined dichotomous and polytomuos item test model in an achievement test. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320.


<https://doi.org/10.21831/pep.v12i2.1433>.

- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus. In ??? (Eds.), *5th ICRIEMS Proceedings* (pp. 479-485). Yogyakarta, Indonesia: Faculty of Mathematics and Natural Sciences. <http://seminar.uny.ac.id/icriems/proceeding2018>.
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://doi.org/10.31327/jme.v6i2.1607>.
- Linacre, J. M. (2012). *Winstep : Rasch-model computer programs*. Chicago: Winsteps.Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset Pemasaran*[Marketing Research]. Jakarta: Eirlangga.
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. In ??? (Eds.), *Current research in Pandemic Covid-19 Era in Indonesia - 2nd ICOSETH 2020* (pp. 44-55). Surakarta, Indonesia: University of Sebelas Maret. <https://doi.org/10.20961/ijsascs.v4i1.49457>.
- Plomp, J. (2013). *Educational design research: An introduction*. Netherlands: Netherlands Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>.
- Prasetya, A., Rosidin, U., & Herlina, K. (2019). Development of instrument sssessment for learning the polytomous response models to train higher order thinking. In ??? (Eds.), *Young Scholar Symposium on Trandisciplinaty in Education and Environment (YSSTEE)* (pp. 1-11). Bandar Lampung: UIN Raden Intan. <https://doi.org/10.1088/1742-6596/1155/1/012032>.
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://doi.org/10.15294/jere.v9i2.46133>.
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Yogyakarta, Indonesia: Nuha Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>.
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran*[Learning evaluation and assessment]. Yogyakarta: Media Akaademi.
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst : Jurnal Pendidikan Fisika dan Terapannya*/Karst : *Journal of Physics Education and Its Application*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>.

- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam questions for Islamic religious education and character at the elementary school level]. *An-Nahdhah/An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article/view/40>.
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Classical tes theory vs response theory? *Didaktika: Jurnal Kependidikan/Didactics: Journal of Education*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>.
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*[Rasch Modeling Applications in Educational Assessment]. Yogyakarta: Trim Komunikata. **need sentence case**
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran*[Development of diagnostic tests in learning]. Yogyakarta: Graha Ilmu.
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>.
- Treagust, D. (1988). Development and use of diagnostic tests to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169.
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep*[Analysis of research data with rash and winstep models]. Bengkulu: Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122-1?inline=1>.
- Widiyatmoko, A., & Shimizu, K. (2018). The Development of two-tier multiple choice test to assess students' conceptual understanding about light and optical instruments. *Jurnal Pendidikan IPA Indonesia*, 7 (4), 491-501. <https://doi.org/10.15294/jpii.v7i4.16591>.
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian*[Using performance assessment for science learning and research]. Bandung: UPI Press.
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>.

2 attachments

 **CORRECTION REPORT 2_Article Sugeng Sutiarto et al.docx**
24K

 **Revision 2_Article Sugeng Sutiarto et al.docx**
1348K

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTIARSO <sugeng.sutiarto@fkip.unila.ac.id>

Mon, Mar 14, 2022 at 3:14 AM

Dear Dr. Sutiarto,

We have received your second revised paper We have sent it to our reviewers again in order to check. We will inform you when we get the result from our reviewers.

If the reviewers confirm your revised paper, we will send the acceptance letter to you.

Thank you for your patience.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

[Quoted text hidden]

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTIARSO <sugeng.sutiarto@fkip.unila.ac.id>

Tue, Mar 15, 2022 at 1:57 PM

In-text citations are not visible in the edited file (we guess it's because of the program you were using). Could you correct the citations and re-send it please urgently?

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

[Quoted text hidden]

SUGENG SUTIARSO <sugeng.sutiarto@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Tue, Mar 15, 2022 at 7:14 PM

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research.

I apologize for my error in citation (there is a problem in my computer). Here I re-send my article.
Thank you for this opportunity to improve.

Best regards,
Sugeng Sutiarto
Lampung University, Indonesia.

[Quoted text hidden]

 **2nd Revision_Article Sugeng Sutiarto et al.docx**
1352K

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTIARSO <sugeng.sutiarto@fkip.unila.ac.id>

Tue, Mar 15, 2022 at 7:57 PM

2nd ROUND CORRECTION REPORT			
No	Reviewer Code	Reviews	Corrections made by the author
1.	R2612	<p>Title: The title should be “The development of an assessment instrument using Polytomous response in mathematics”</p>	<p>Title: The Development of an Assessment Instrument Using Polytomous Response in Mathematics</p>
2.	R2612	<p>Abstract: Please use past tense in the abstract. For example, write showed instead of show. Use “Questionnaire was to identify” instead of “Questionnaire is to identify”</p>	<p>Abstract: The study was conducted in three vocational schools in Lampung, Indonesia. The study involves 413 students consisting of 191 male students and 222 female students. The data were collected through questionnaires and tests. Questionnaire was to identify instruments commonly used by teachers so far and to validate instruments by experts. The test used multiple-choice tests with open reasons as many as 40 items. The data were analyzed in two ways, namely analysis with classical and modern theories. (p.1)</p>
3.	R2612	<p>Introduction: The revisions to my previous comments are not satisfactory. The introduction was shortened but its current version does not still explain what the problem is. From the authors’ writing, we do not have adequate information about the strengths and weaknesses of the existing instruments. The explanations are superficial.</p> <p>My previous questions remain valid. What do we know about the existing instruments? What do we need to know about the existing instruments?</p>	<p>Introduction: What do we know about the existing instruments? Each type of test has strengths or weaknesses with each other. The strength of multiple-choice tests over essays is that multiple-choice tests can be conducted for many students, are more objective, and the test results can be known more quickly; but has a weakness, namely the multiple choice test is not able to see the actual abilities of students and the answers tend to guess or try it out (Rosidin, 2017). In addition, the strength of the multiple choice test has a scoring certainty compared to the essay test, namely 1 and 0 (Getting a score of 1 for the correct answer, and a score of 0 for the wrong answer choice). (p. 2)</p> <p>researchers have found weaknesses in the test, such as students' misconceptions cannot be known in detail (Antara, et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), student answers are still guessing (Myanda, et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors is easy to observe (Treagust, 1988), and The suitability between the student's answer choices and the reason is easy to know (Diani, et al., 2019). (pp. 3-4)</p>

			<p>However, this classical theory has a weakness, namely it cannot separate the characteristics of students and items, also the characteristics of items will change when students change. So, classical theory is considered less able to provide information about students' actual abilities. (p. 5)</p> <p>What do we need to know about the existing instruments? To reduce the weakness of the closed polytomous response test, the experts modified the test to be the open polytomous response test. (p. 4)</p> <p>Until now, research on the open polytomous response test is still very limited, (p. 4)</p> <p>Because there is still limited research on the open polytomous response test, it is necessary to conduct research on research subjects with other characteristics. This research was conducted on students in vocational schools who have different characteristics from students in college or high school. The characteristic difference is students in vocational schools place mathematics as a secondary subject, while students in colleges and high schools place mathematics as a primary subject (Oktaria, 2016). In addition, graduate students in vocational schools are more oriented towards practical abilities and skills, and in contrast to graduate students in colleges and high schools are academically oriented (Permendikbud, 2016). Therefore, it is reasonable to conduct research by developing the open polytomous response test for students in vocational schools. The test instrument developed must be accountable as a condition of a good test, and it is necessary to analyze the quality of the item (Rosidin, 2017). (p. 4)</p> <p>Therefore, experts suggest that the test instrument is accountable, the quality of the items must be good according to the analysis of classical and modern theory (p. 5)</p>
--	--	--	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

4.	R2612	<p>Method: Use the past tense “The instrument is validated by two people...”</p>	<p>Method: The instrument was validated by two people</p>
5.	R2612	<p>Discussion: I could not understand “The similarity of Sarea's research is safe...”</p> <p>The word “the product” is not acceptable in an educational paper. - “Although the products are different ...”.</p>	<p>Discussion: “The similarity of Sarea's research is safe...” [The word "safe" removed] The similarity with Sarea's research lies in the research objectives and the analysis used (classic and modern). (p. 26).</p> <p>“the product” is not acceptable in an educational paper. - “Although the products are different ...”.</p> <p>[The word "product" is removed, and the sentence is corrected] Likewise with Saepuzaman's research the closed polytomous response test provide confidence that items that are not good according to classical theory are actually good items according to modern theory. The results of previous studies have provided support for the development of instruments on the polytomus response test, ...</p>
6..	R2612	<p>Conclusion: The new knowledge is not explained in the literature?</p> <p>Rewrite – “the condition that must exist so that an instrument with a response polytomous can be used to determine students' abilities is the suitability of the results of the analysis between classical and modern theories”</p>	<p>Conclusion: [The sentence at the conclusion is changed] So, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory) (p. 27)</p> <p>It is based on the literature: Therefore, experts suggest that the test instrument is accountable, the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014). (p. 5)</p>
7.	R2612	<p>Language: The use of the language is still problematic.</p>	<p>Language: Overall, the sentence in the article has been improved.</p>

8.	R2612	Test: The developed instrument should be given in an appendix.	Test: The developed instrument is already in the appendix. Appendix Instrument of the Open Polytomous Response Test (pp. 33-36)
----	--------------	--------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------

The Development of an Assessment Instrument Using Polytomous Response in Mathematics

Abstract: This research is development research aimed to produce a good instrument of assessment in mathematics using polytomous response according to classical and modern theories. This research design uses the Plomp model which consists of five stages, namely: preliminary investigation, design, realization or construction, trial, revision, and implementation (testing). The study was conducted in three vocational schools in Lampung, Indonesia. The study involves 413 students consisting of 191 male students and 222 female students. The data were collected through questionnaires and tests. Questionnaire was to identify instruments commonly used by teachers so far and to validate instruments by experts. The test used multiple-choice tests with open reasons as many as 40 items. The data were analyzed in two ways, namely analysis with classical and modern theories. The results show that the open polytomous response test have a good category according to classical and modern theory, and the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Key words: assessment for learning, classical and modern theory, multiple choice tests with open reason, polytomous response, vocational school

Introduction

Assessment is an important activity and needs to be done by teachers in schools. The conventional paradigm often interprets the assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Sya20). If referring to the current paradigm, assessment in schools is divided into three parts, namely assessment as learning, assessment for learning, and assessment of learning (Wul18). Assessment as learning has almost the same function as assessment for learning but assessment as learning involves students in assessment, such as

assessing themselves or colleagues. Assessment of learning is an assessment carried out after all learning ends and aims to assess student achievement. Assessment for learning is an assessment carried out during the learning process and aims to improve learning. It can play a role in preventing students from experiencing further learning failure because of its position between the other two assessments (Ear13) as shown in the following assessment pyramid (Figure 1).

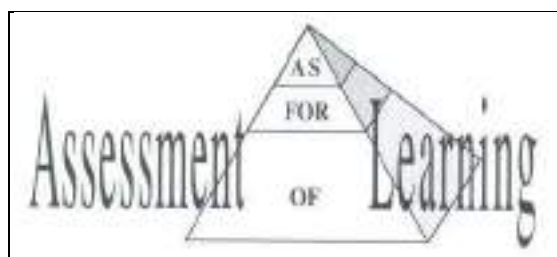


Figure 1. Assessement Pyramid

Assessment activities can be applied with tests. A test is a tool or procedure used to find out or measure students' abilities about something with certain rules (Ari12). A test consists of two types, namely multiple choice and essay. Multiple choice test is a form of assessment in which each item provides an answer choice, and one of the choices is the correct answer. The essay test is a form of assessment that requires answers in sentences or words. Each type of test has strengths or weaknesses with each other. The strength of multiple-choice tests over essays is that multiple-choice tests can be conducted for many students, are more objective, and the test results can be known more quickly; but has a weakness, namely the multiple choice test is not able to see the actual abilities of students and the answers tend to guess or try it out (Ros17). In addition, the strength of the multiple choice test has a scoring certainty compared to the essay test, namely 1 and 0 (Getting a score of 1 for the correct answer, and a score of 0 for the wrong answer choice). Multiple choice tests with only two answer choices are called dichotomous tests, and multiple choice tests with more than two answer choices are called polytomous tests (Kar08).

Until now, multiple choice tests are still widely used by teachers to assess students' abilities, especially students with a large number and wide area. To reduce the weakness of multiple choice tests, in the last four decades, experts have developed multiple choice tests by combining multiple choice tests and essays into multiple choice tests with reasons, and called polytomous tests with responses; or abbreviated the polytomous response test (Suw12). The polytomous response test score is 1-4. Score 4 for the correct answer and reason, score 3 for correct answer but wrong reason, score 2 for wrong answer but correct reason, and score 1 for wrong answer and reason (Kar08).

In the 80s, the first time that was focused on and developed by experts was the closed polytomous response test (Tre88), and this test aims to diagnose misconceptions in biology, physics, and chemistry. This test consists of two levels; The first level is choosing answers on the multiple choice test, and the second level is choosing reasons based on the answer choices at the first level (Cha07). Several studies on the closed polytomous response test have been carried out, such as tests on light and optical materials (Wid18), test on calculus material (Khi18), test on acid and base material (And20), test on reasoning material (Placeholder2), test on human digestive system material (Jam21), test on mathematical connection material (Les21). Meanwhile, the development of multiple-choice tests with open reasons is still limited, namely mathematics in calculus (Yan17), and outside mathematics, namely physics in Higher Order Thinking Skills (Pra19). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions cannot be known in detail (Ant19), the test instrument is difficult to construct (MKh19), student answers are still guessing (Mya20). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors is easy to observe (Tre88), and The suitability between the student's answer choices and the reason is easy to know (Dia19).

To reduce the weakness of the closed polytomous response test, the experts modified the test to be the open polytomous response test. The open polytomous response test is a form of multiple choice test that provides a place to write arguments for the answer choices (Ret14). Until now, research on the open polytomous response test is still very limited, such as tests on calculus material in universities (Yan17), and tests on physics material in high school (Pra19). Because there is still limited research on the open polytomous response test, it is necessary to conduct research on research subjects with other characteristics. This research was conducted on students in vocational schools who have different characteristics from students in college or high school. The characteristic difference is students in vocational schools place mathematics as a secondary subject, while students in colleges and high schools place mathematics as a primary subject (Okt16). In addition, graduate students in vocational schools are more oriented towards practical abilities and skills, and in contrast to graduate students in colleges and high schools are academically oriented (Per162). Therefore, it is reasonable to conduct research by developing the open polytomous response test for students in vocational schools. The test instrument developed must be accountable as a condition of a good test, and it is necessary to analyze the quality of the item (Ros17).

There are two theories in analyzing item quality, namely classical and modern theory. Classical theory is a measurement theory for assessing tests based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Ham93). Modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and is known as Item Response Theory/IRT (Ham82). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely it cannot separate the characteristics of students and items, also the characteristics of items will change when students change. So, classical

theory is considered less able to provide information about students' actual abilities. Modern theory is a solution to overcome the weaknesses of classical theory, because in modern theory an item does not affect other items (local independence), and items only measure one dimension/unidimensional (Ani13) , and an item eliminates the relationship between respondents and items (parameter invariance) (Sae21). Therefore, experts suggest that the test instrument is accountable, the quality of the items must be good according to the analysis of classical and modern theory (Ret14).

This research is a development research that aims to produce a good open polytomous response test according to classical and modern theory. The problem formulations proposed are (1) does the open polytomous response test have a good category according to classical and modern theory?, and (2) can the open polytomous response test provide information on the actual competence of students?.

Methodology

Research Design

This research is a research and development that refers to Plomp's model (Plo13) with the research procedure consisting of five stages, namely preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

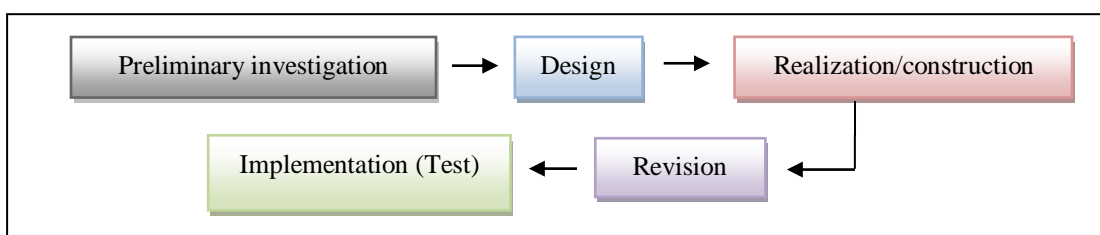


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the assessment instruments used by teachers so far. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make an expert assessment questionnaire sheet. The realization or construction stage is developing the items test, and also the expert validation process for the items test. The revision stage is the improvement of items test based on expert advice. The implementation (test) stage is to try out the test to students, and analyze the results of the test.

Research Subject

The subjects of the study are students of a vocational school in the province of Lampung, Indonesia. The research sample is determined using a non-probability sampling technique in the form of accidental sampling. It means taking the subject based on a subject that is easy to find and ready to be a respondent (Mal06). The selected schools are three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects are 413 students at grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 from the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38

	I-G	40	8	23	67.74
SMK Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collecting Technique

Data were collected using a questionnaire and test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed and to determine content validity (Suh21). **The instrument was validated by two people** who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert are the suitability of the items with the indicators, language, and alternative answers to the questions. The score in validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining content validity, the instrument was tested on students. Then, it was continued by determining the validity of the construct and its reliability, with the aim that the instrument can be further analyzed.

The instrument used was the open polytomous response test which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contains five answer choices along with the reasons. Student answers score refers to the polytomous score in the Partial Credit Model, where answer choices and reasons are related (Ret14), as shown in Table 3 below.

Table 3. Student Answers Score

Student Answers	Score
------------------------	--------------

Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The research data obtained are analyzed in two stages, namely (1) questionnaire data analysis (qualitative analysis), and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis.

1. Questionnaire data analysis (qualitative analysis)

The questionnaire data analyzed include two parts, namely identification data on the instruments used by the teacher and expert validation data. The identification data on the instrument are analyzed descriptively, and the expert validation data are analyzed for trends or expert agreement using the Gregory formula (Gre15), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range 0 - 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the value of the validity of an item.

2. Test data analysis (empirical analysis)

After the content validity test, it is followed by construct validity and reliability tests. The construct validation test uses exploratory factor analysis. The instrument is considered to

have a good construct if the explained Kaiser-Meyer-Olkin (KMO) value is more than 0.5 (Ret14). Reliability test using Cronbach's Alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Ari12). If the instrument has good construct validation, further tests can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is because they are both preliminary analyzes of the assumptions of measurement theory (Ham93). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program is used for classical theory, and the Winsteps program for modern theory (Sar19). The Winsteps program is used because it has several advantages (Unt20), namely, it can analyze polytomous data and can analyze the maximum likelihood model using a 1-parameter logistic model.

2.1 Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answer the questions correctly or incorrectly. The difficulty of a good (medium) item is if the index is in the range of 0.3 to 0.7. If the index is below 0.3 then the item is difficult, and vice versa if the index is above 0.7 then the item is easy.
- b. Item discrimination is the item's ability to distinguish high-ability students from stupid, low-ability students. Good item discrimination if it has an index above 0.3; and if the item discrimination index is below 0.3 then the item needs to be revised (Ari12).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items identifies the ability of about 50% of respondents who are expected to answer items correctly (DeM10). An item is said to be good if it has an index between -2 and +2 (Ham85). If the index is close to -2, then the item

is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Ret14). In the Winsteps program, the item difficulty level is in the Measure column.

- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Cro86). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column.

According to modern theory, before analyzing the level of difficulty and item discrimination, three assumptions must be tested, namely unidimensional, local independence, and model fit (Ham91). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigen values of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals on the unidimensional solution (DeM10). In this study, the dimensional test uses the Eigen value analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer on each item. If the unidimensional assumption has been met, the local independence assumption has also been met (DeM10). Model fit test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring Outfit Mean Square (MNSQ) and Pt-Measure. If the Outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Lin12). In addition, the information function and standard error

measurement (SEM) are analyzed which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it is found that so far the teacher has never used the polytomous response instrument with a multiple-choice test with open reasons. As many as 80% of teachers use essay tests and 20% of teachers use multiple choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers use this assessment as a learning improvement, such as improving lesson plans and teaching models/methods. The results of the questionnaire stated that 90% of teachers who do not use assessment as an improvement in learning are caused by several aspects, such as teachers are not understanding assessment (20%), teachers are not knowing how to analyze assessments (50%), and teachers are not knowing how to develop good assessment questions (30%). The following is a summary of the questionnaire data from the identification of assessment for learning instruments.

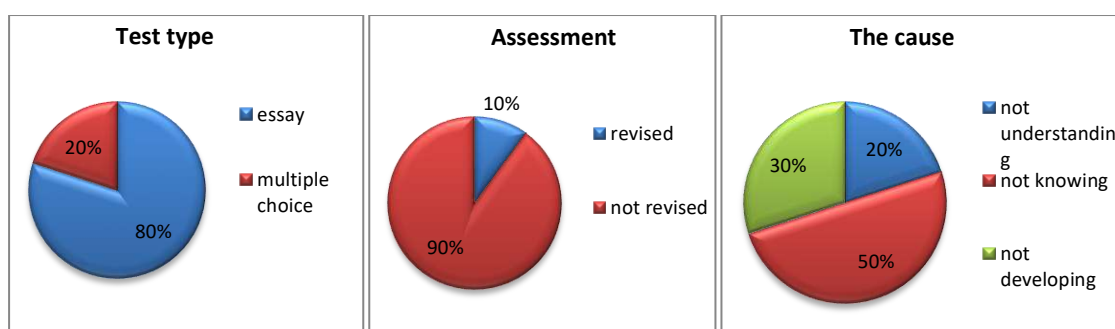


Figure 3. Description of Teacher Condition in Assessment for Learning

Content Validity

The results of the two expert assessments show that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement with the Gregory Index formula is

obtained as shown in Table 4 below.

Table 4. Index Gregory Items

		Rater 1	
		weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provide some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arrange them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it is followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained the Cronbach's

Alpha coefficient value of 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to the classical and modern methods.

Table 6. Item Reliability

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data with classical theory does not require testing assumptions, but the analysis of the level of difficulty and item discrimination can be directly calculated. The results of the analysis of the level of difficulty and item discrimination are obtained as shown in Table 7 below.

Table 7. Level of Difficulty and Item Discrimination

Item	Level of difficulty	Category	Discrimination	Category	Item	Level of difficulty	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised
4	0.540	good	0.304	Good	24	0.438	good	-0.071	revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	revised

Based on Table 7, it is found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are included in the good category. Meanwhile, only two items on

discrimination have good categories, and the remaining items need to be revised. The results indicate that all items are good based on the level of difficulty, but almost all items need to be revised for item discrimination.

Analysis of Test Data with Modern Theory

Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, then calculating the eigenvalues. The eigenvalue is then used to calculate the percentage of explained variance, as well as describe the scree plot (Ret14). The output of factor analysis is the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is based on the cumulative percentage of eigenvalues and scree plots. If the cumulative percentage of eigenvalues in the first factor is more than 20%, then the unidimensional assumption is accepted (Ret14). In Table 9, it can be seen that the cumulative percentage of the eigenvalues in the first factor is 20.220%. The cumulative percentage of the eigenvalues has exceeded 20%, so the instrument in the study is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalues		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot which is based on the

number of factors marked by the steepness of the graph with the acquisition of eigenvalues.

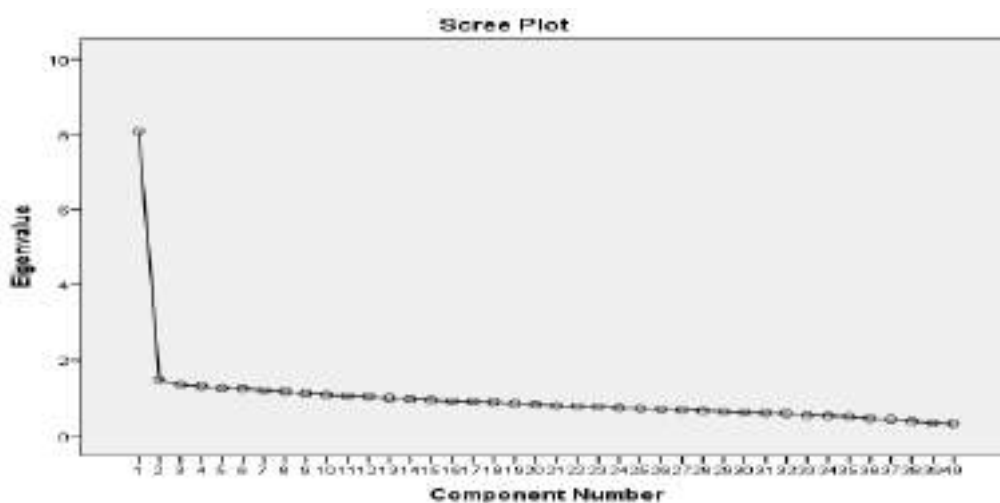


Figure 4. Scree Plot Unidimensi

Based on the scree plot, it is known that the eigenvalues immediately slope on the second factor. It shows that there is only one dominant factor in the developed instrument. The results prove that the test kit meets the unidimensional assumption or in other words only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be accepted if the respondent's answer to one item does not affect the respondent's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. It confirms that the assumption is automatically proven after being proven by the unidimensionality of the respondent's data on a test (Ret14).

Table 10. Covariance Matrix

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	

K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196
------------	-------	--------	--------	--------	--------	--------	--------	--------	--------	--------

Data in Table 10 shows the results of the variance-covariance values between groups of students' abilities. It can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. It can be concluded that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test is analyzed using the Winsteps program. The item requirements are called fit to the model if the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Corr is positive, then it can be said that the item fits the model (Sum15). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items match the model or fit (Table 11).

Table 11. Item Fit on Model

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73
 Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93
 Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Difficulty Level

The item difficulty level is analyzed using the Winsteps program, and the results obtained can be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of -2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sum15). It can also be seen on the difficult map items, namely the difficulty level is in the range of -2 and 2.

Table 12. Item Difficulty Level

The item discrimination analysis used the Winsteps program and the results are presented in Table 13 (column PT-Measure Corr). Item discrimination is in the range of 0.23 to 0.74; with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Placeholder1).

Table 13. Item Discrimination

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73

Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Comparative Analysis between Classical and Modern Theory

The results of the analysis of classical and modern theory obtained the index of difficulty level and item discrimination as follows.

Table 14. Analysis Classical and Modern Theories

Parameter	Classical Theory	Modern Theory
-----------	------------------	---------------

	Many Items with Good Category	Percentage	Many Items with Good Category	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 14, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in good category than using classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Table 13), it can be seen that there is a match between the categories of item discrimination. It means that if the item discrimination is not good with the classical theory, then the item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability is measured by using a test that is expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error or Standard Error Measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

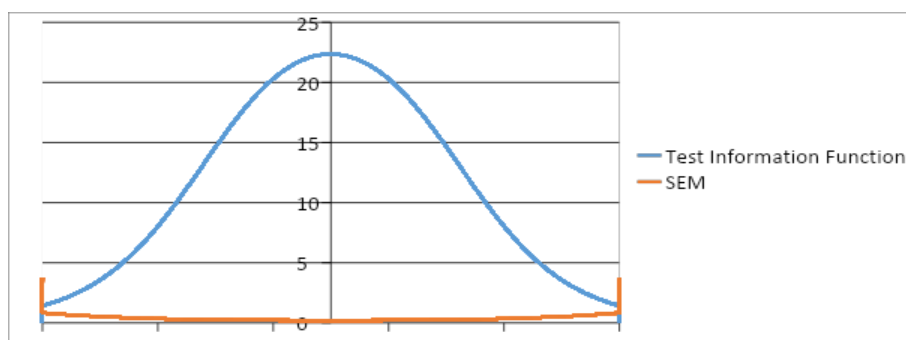


Figure 5. Graph of Information Function and Measurement Error

Figure 5 shows that the instrument provides a maximum of 22.36 information and has the smallest measurement error of 0.21 if it is given to students with moderate ability, which is 0.2. The lower limit and upper limit of the interval is the ability score where the graph of the information function and the SEM graph intersect in that interval. The graph indicates that the greater the value of the information function, the smaller the measurement error (SEM). Item information function states the strength or contribution of test items in revealing the latent trait as measured by the test. With the item information function, it is known which items match the model, thus helping in the selection of test items (Ret14). In conclusion, the characteristics of the test kit are suitable for students with moderate abilities.

Based on the test results, in addition to knowing the quality of the developed test instruments, it can also be seen the ability of students to work on the questions given. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Blo56). A total of 6 student answers were selected as samples with different abilities (high, medium, and low).

Item 1: the cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms. arithmetic sequence.

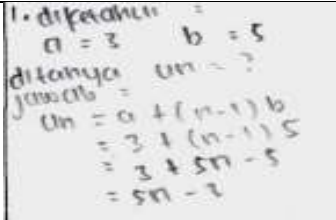
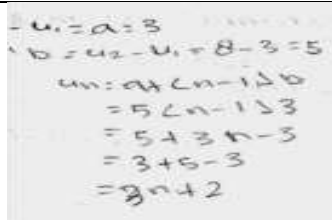
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	 <p>1. dipecahkan = $a = 3$ $b = 5$ ditanya $U_n = ?$ jawab = $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	 <p>$U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>

Figure 6. Student Answers in Item 1

Item 2: the cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand and determine the number of terms in a sequence by using the general formula for an arithmetic sequence or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

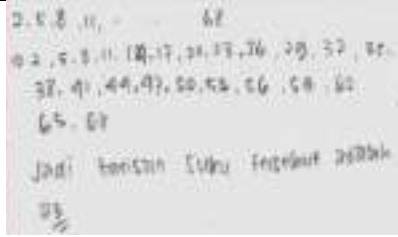
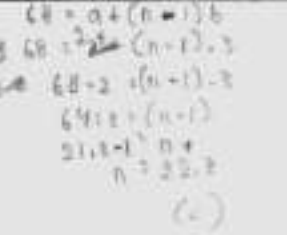
Question 2:	Pattern 1	Pattern 2
<p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>		

Figure 7. Student Answers in Item 2

Item 3: the cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and determined the difference or difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

Question 3:	Pattern 1	Pattern 2
-------------	-----------	-----------

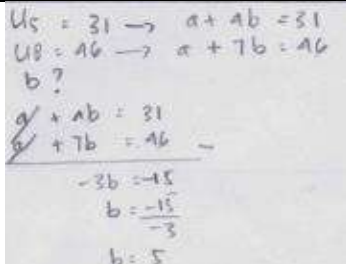
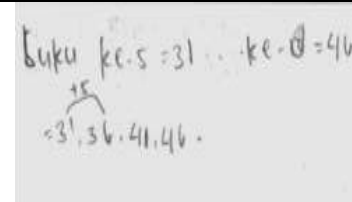
<p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p>	 <p> $U_5 = 31 \rightarrow a + 4b = 31$ $U_8 = 46 \rightarrow a + 7b = 46$ $b = ?$ $a + 4b = 31$ $+ 7b = 46$ $-3b = -15$ $b = \frac{-15}{-3}$ $b = 5$ </p>	 <p> buku ke-5 = 31 .. ke-8 = 46 $+5$ $= 31, 36, 41, 46$ </p>
<p>Reason:</p>		

Figure 8. Student Answers in Item 3

Item 4: the cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula, but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

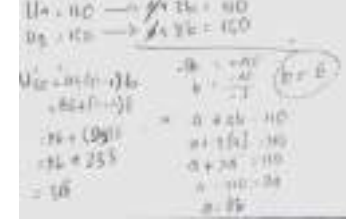
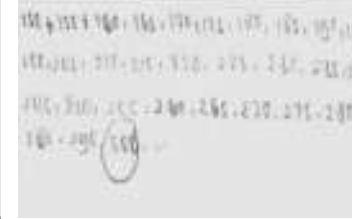
Question 4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p>	 <p> $U_4 = 110 \rightarrow a + 3b = 110$ $U_9 = 150 \rightarrow a + 8b = 150$ $-5b = -40$ $b = 8$ $a + 3(8) = 110$ $a + 24 = 110$ $a = 110 - 24$ $a = 86$ </p>	 <p> $110 + 114 + 118 + 122 + 126 + 130 + 134 + 138 + 142 + 146 + 150$ $154 + 158 + 162 + 166 + 170 + 174 + 178 + 182 + 186 + 190 + 194 + 198 + 202 + 206 + 210 + 214 + 218 + 222 + 226 + 230 + 234 + 238 + 242 + 246 + 250 + 254 + 258 + 262 + 266 + 270 + 274 + 278 + 282 + 286 + 290 + 294 + 298 + 302 + 306 + 310 + 314 + 318 + 322 + 326 + 330 + 334 + 338 + 342 + 346 + 350$ </p>
<p>Reason:</p>		

Figure 9. Student Answers in Item 4

Item 5: the cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but by writing the terms from known terms and inserts several terms and then defines them.

Question 5:	Answer Pattern 1	Answer Pattern 2
-------------	------------------	------------------

An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
Then the middle term of the sequence is ...308

A. 53
B. 52
C. 20
D. 11
E. 10

Reason:



Figure 10. Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test and all parameters have been accepted. This instrument is a combination of multiple choice test and essay. Multiple choice tests are easier to check students' answers but students' mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find out deeper mathematical thinking processes but it takes a long time to check the answers.

Assessment of learning instruments that have been carried out with item analysis and the results of student analysis is one of the important sources of composite scores to be reported. In the final report, the test taker's ability score should be changed to a score of 0 - 10 from 0 - 100, according to the needs of the school. The transformation uses a linear transformation by dividing the score by the ideal score and then the result is multiplied by 10 to get a value in the range 0 - 10 or multiplied by 100 to get a score of 0 - 100. In the range 0 - 10, the score obtained by students taking the test the highest mathematics learning was 8.56 and the lowest was 4.31. In the range 0 - 100, the scores obtained by students with the highest mathematics is 85.625 and the lowest is 43.125.

The results of the assessment of students' mathematical abilities are presented in the form of very low to very high predicates. The results of the test analysis show that most students have low and very low abilities, namely 62% (253 students). Meanwhile, students who have high

and very high abilities are 38% (160 students). The results of another analysis find that students who have high abilities tend to work according to the concepts that have been given by the teacher but do not follow the completion steps, students who have moderate abilities can solve problems according to the concepts that have been given by the teacher and the steps, and there are students who have abilities but are not able to use the concepts given by the teacher and are not even able to give clear reasons..

Another result of this study is that teachers agree to provide learning assessments with multiple choice questions with open-ended reasons because the instrument is easier for teachers to find out students' difficulties in certain materials. In this way, teachers can also provide remedial or other assistance to students who have learning difficulties. It means that the polytomous response instrument can be used as a way to determine which students need remedial or not. In general, previous research states how to determine students who need remedial only one test, namely multiple-choice tests (Gie17) or essays (Ana20).

Discussion

This research is development research to produce the open polytomous response test. The instrument is a multiple choice test with open reasons. This instrument is analyzed by classical and modern theory. There are differences in the results of the analysis between classical and modern theories, namely item discrimination. Classical theory analysis obtains 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtains 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as not good but the results of the modern analysis are categorized as good, and vice versa (Ret14). It means, if you find items that are not in a good category with classical theory, do not be in a hurry to revise or replace

them before the analysis of modern theory analysis.

Research on assessment for learning with the open polytomous response test is still limited. When compared to previous research, there is only one study on assessment for learning with polytomous responses (Yan17). However, Yang study has several fundamental differences, namely the research objectives and data analysis. The research aims to diagnose student errors in university on the concept of calculus, not to produce a good assessment instrument. The data analysis uses parametric statistics (covariance), not using item analysis (classical and modern). Since the objectives and data analysis are different, the results of the study cannot be compared with the results of this study. However, this research has provided a reference for researchers in making reasoned multiple choice tests, such as the suitability of items with indicators, language, and alternative answers to questions.

Other studies are similar to assessment for learning with the open polytomous response test (Sar18). The similarity with Sarea's research lies in the research objectives and the analysis used (classic and modern). However, the difference is the researchers do not develop their questions and the questions are in the form of the closed polytomous response test. The results of Sarea's research states that the comparison of the results of the classical and modern theory of item analysis is different. The difference is that the level of difficulty and item discrimination in the classical theory is more categorized as good than the modern method. In other words, the modern way of stating the level of difficulty and item discrimination is categorized as good even though the analysis method states that the items are categorized as not good. Likewise with Saepuzaman's research the closed polytomous response test provide confidence that items that are not good according to classical theory are actually good items according to modern theory. The results of previous studies have provided support for the development of instruments on the polytomus response test, and this test instrument can be

used as an alternative to all learning assessment (assessment: as learning, for learning, and of learning) for all vocational schools in Lampung, Indonesia and even outside Indonesia.

Conclusion

Based on the results of the research and discussion, conclusions are obtained, namely (1) the open polytomous response test have a good category according to classical and modern theory. So, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. this is observed in the students' arguments in giving reasons for their choices. Therefore, the open polytomous response test can can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, teachers should familiarize students with giving a test in the form of a polytomous response before giving the test. For schools, principals or other leaders should encourage other teachers to take advantage of this test, and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important, so that students' prior knowledge can be known so that learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not been the researchers' expectations, for example representing schools with high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic material (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Ambarwati, R., Sunardi, Yudianto, E., Murtikusuma, R. P., & Safrida, L. N. (2020). Developing mathematical reasoning problems type two-tier multiple choice for junior high school students based on ethnomathematics of jember fashion carnival. In Suratno (Ed.) *ICOLSSSTEM*. IOP Publishing. <https://doi.org/10.1088/1742-6596/1563/1/012036>.
- Andaria, M. & Hadiwinarto. (2020). Development of a two-tier multiple choice question assessment instrument to measure students science process skills on acid-base material. *ISEJ: Indonesian Science Education Journal*, 1 (3), 257-268. <https://siducat.org/index.php/isej/article/view/141>.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi/ Journal of Mathematics, Statistics, & Computing*, 9(2), 95-113. <https://doi.org/10.20956/jmsk.v9i2.3402>.
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *Jurnal Suluh Pendidikan/Journal of Educational Suluh*, 17(1), 32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.

- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: the cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-Tier Diagnostic Test With Certainty of Response Index on The Concepts of Fluid. In C. Anwar (Ed.), *Young Scholar Symposium on Transdisciplinary in Education and Environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf>
- DeMars, C. E. (2010). *Item response theory*. Oxford University Press.
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Hong Kong: Assessment & Support Team. <https://wlts.edb.hkedcity.net/en/home/AandLI2.html>.
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87 (6), 1082–1116. <https://doi.org/10.3102/0034654317726529>.
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12 (3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543>.
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Jamhari, M. (2021). Developing the two-tier multiple choice tests in enhancing students' higher-order thinking skills on human digestive system. *Eduproxima: Jurnal Ilmiah Pendidikan IPA*, 3 (1), 50-64. <https://doi.org/10.29100/eduproxima.v3i1.1853>.
- Kartono. (2008). Equating the combined dichotomous and polytomous item test model in an achievement test. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>.

- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how?. In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <http://seminar.uny.ac.id/icriems/proceeding2018>.
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://doi.org/10.31327/jme.v6i2.1607>.
- Linacre, J. M. (2012). *Winstep : Rasch-model computer programs*. Winsteps. Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset Pemasaran* [Marketing Research]. Eirlangga.
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. In Cari (Ed.), *Current research in Pandemic Covid-19 Era in Indonesia - 2nd ICOSETH 2020* (pp. 44-55). University of Sebelas Maret. <https://jurnal.uns.ac.id/ijsacs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Content standards for primary and secondary education]. Indonesian Government publication service.
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>.
- Prasetya, A., Rosidin, U., & Herlina, K. (2019). Development of instrument assessment for learning the polytomous response models to train higher order thinking. In C. Anwar (Ed.), *Young Scholar Symposium on Transdisciplinary in Education and Environment (YSSTEE)* (pp. 1-11). UIN Raden Intan. <http://repository.lppm.unila.ac.id/11982/1/33.%20Prasetya%202019.J.Phys.Conf.Ser..pdf>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://doi.org/10.15294/jere.v9i2.46133>.
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>.

- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst : Jurnal Pendidikan Fisika dan Terapannya/ Karst : Journal of Physics Education and Its Application*, 4 (1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>.
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam questions for Islamic religious education and character at the elementary school level]. *An-Nahdhah/An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article/view/40>.
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Classical tes theory vs response theory? *Didaktika: Jurnal Kependidikan/Didactics: Journal of Education*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>.
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunitas.
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic tests in learning]. Graha Ilmu.
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9 (4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>.
- Treagust, D. (1988). Development and use of diagnostic tests to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169.
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep* [Analysis of research data with rash and winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122-1?inline=1>.
- Widiyatmoko, A., & Shimizu, K. (2018). The Development of two-tier multiple choice test to assess students' conceptual understanding about light and optical instruments. *Jurnal Pendidikan IPA Indonesia*, 7 (4), 491-501. <https://doi.org/10.15294/jpii.v7i4.16591>.
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive

mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>.

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :

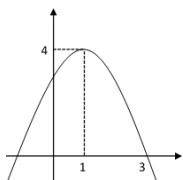
Class/Department :

School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

<p>1. Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ C. $U_n = 4n - 1$ B. $U_n = 5n - 2$ D. $U_n = 3n + 2$ C. $U_n = 4n - 1$</p> <p>Reason:</p>	<p>2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 D. 23 B. 13 E. 24 C. 22</p> <p>Reason:</p>
<p>3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 D. 8 B. 6 E. 11 C. 7</p> <p>Reason:</p>	<p>4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 D. 344 B. 318 E. 354 C. 326</p> <p>Reason:</p>
<p>5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...</p> <p>A. 53 D. 11 B. 52 E. 10 C. 20 D. 11 E. 10</p> <p>Reason:</p>	<p>6. Given the arithmetic sequence: 4, 10, 16, 22, If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...</p> <p>A. 18 D. 24 B. 20 E. 26 C. 22</p> <p>Reason:</p>
<p>7. The nth term of an arithmetic series is $U_n = 3n - 5$. The formula for the sum of the first n terms of the series is ...</p> <p>A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$ B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$ C. $S_n = \frac{n}{2}(3n - 4)$</p> <p>Reason:</p>	<p>8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....</p> <p>A. $5n - 20$ D. $2n - 20$ B. $5n - 10$ E. $2n - 10$ C. $2n - 30$</p> <p>Reason:</p>
<p>9. The sum of all integers between 100 and 300 which are divisible by 5 is ...$S_n = \frac{n}{2}(3n - 7)$</p> <p>A. 8.200 D. 7.600 B. 8.000 E. 7.400 C. 7.800</p> <p>Reason:</p>	<p>10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?</p> <p>A. 24 D. 27 B. 25 E. 28 C. 26</p> <p>Reason:</p>

<p>23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah</p> <p>...</p> <p>A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$</p> <p>B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$</p> <p>C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$</p> <p>Reason:</p>	<p>24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.</p> <p>If $A + B = C$, then $x + y = \dots$</p> <p>A. -5 D. 3</p> <p>B. -1 E. 5</p> <p>C. 1</p> <p>Reason:</p>
<p>25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah</p> <p>...</p> <p>A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$</p> <p>B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$</p> <p>C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$</p> <p>Reason:</p>	<p>26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$</p> <p>then $A(B - C) = \dots$</p> <p>A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$</p> <p>B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$</p> <p>C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$</p> <p>Reason:</p>
<p>27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...</p> <p>A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$</p> <p>B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$</p> <p>C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$</p> <p>Reason:</p>	<p>28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix. Value of x that satisfies is ...</p> <p>A. -5 D. 3</p> <p>B. -4 E. 4</p> <p>C. -3</p> <p>Reason:</p>
<p>29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ...</p> <p>A. 0 D. 2</p> <p>B. 1 E. 4</p> <p>C. 2</p> <p>Reason:</p>	<p>30. Transpose matrix P adalah P^T. If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ...</p> <p>A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$</p> <p>B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$</p> <p>C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$</p> <p>Reason:</p>
<p>31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$.</p> <p>Inverse matrix AB adalah $(AB)^{-1} = \dots$</p> <p>A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$</p> <p>B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$</p> <p>C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$</p> <p>Reason:</p>	<p>32. The roots of the quadratic equation $3x^2 - 4x - 12 = 0$ are ...</p> <p>A. $x^2 + x - 12 = 0$</p> <p>B. $x^2 - x - 12 = 0$</p> <p>C. $x^2 - x + 12 = 0$</p> <p>D. $x^2 - 3x + 4 = 0$</p> <p>E. $x^2 - 4x + 3 = 0$</p> <p>Reason:</p>
<p>33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...</p>	<p>34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2. If $x_1 > x_2$, then $x_1 - x_2$ is ...</p>

<p>A. $-2 \operatorname{dan} \frac{5}{6}$ B. $2 \operatorname{dan} -\frac{5}{6}$ C. $2 \operatorname{dan} \frac{6}{5}$ Reason:</p>	<p>D. $-2 \operatorname{dan} -\frac{6}{5}$ E. $-2 \operatorname{dan} \frac{6}{5}$ A. -4 B. -2 C. 0 D. 2 E. 4 Reason:</p>
<p>35. The roots of the quadratic equation $2x^2 - 3x - 9 = 0$ are x_1 dan x_2. Value of $x_1^2 + x_2^2$ is ... A. $11\frac{1}{4}$ B. $6\frac{3}{4}$ C. $2\frac{1}{4}$ Reason:</p>	<p>36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α dan β. The quadratic equation $(\alpha - 2)$ dan $(\beta - 2)$ is ... a. $x^2 + 6x + 5 = 0$ b. $x^2 + 6x + 7 = 0$ c. $x^2 + 6x + 11 = 0$ d. $x^2 - 2x + 3 = 0$ e. $x^2 + 2x + 11 = 0$ Reason:</p>
<p>37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ... A. $y = x^2 - 2x + 1$ B. $y = x^2 - 2x + 3$ C. $y = x^2 + 2x - 1$ D. $y = x^2 + 2x + 1$ E. $y = x^2 + 2x + 3$ Reason:</p>	<p>38. The figure below is a graph of the quadratic equation ? ... A. $y = x^2 + 2x + 3$ B. $y = x^2 - 2x - 3$ C. $y = -x^2 + 2x - 3$ D. $y = -x^2 - 2x + 3$ E. $y = -x^2 + 2x + 3$ Reason: </p>
<p>39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0, -4)$ then the value of $f(7)$ is ... A. -16 B. -17 C. -18 D. -19 E. -20 Reason:</p>	<p>40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ... A. $y = -x^2 + 2x - 3$ B. $y = -x^2 + 2x + 3$ C. $y = -x^2 - 2x + 3$ D. $y = -x^2 - 2x - 5$ E. $y = -x^2 - 2x + 5$ Reason:</p>

The Development of an Assessment Instrument Using Polytomous Response in Mathematics

Abstract: This research is development research aimed to produce a good instrument of assessment in mathematics using polytomous response according to classical and modern theories. This research design uses the Plomp model which consists of five stages, namely: preliminary investigation, design, realization or construction, trial, revision, and implementation (testing). The study was conducted in three vocational schools in Lampung, Indonesia. The study involves 413 students consisting of 191 male students and 222 female students. The data were collected through questionnaires and tests. Questionnaire was to identify instruments commonly used by teachers so far and to validate instruments by experts. The test used multiple-choice tests with open reasons as many as 40 items. The data were analyzed in two ways, namely analysis with classical and modern theories. The results show that the open polytomous response test have a good category according to classical and modern theory, and the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Key words: assessment for learning, classical and modern theory, multiple choice tests with open reason, polytomous response, vocational school

Introduction

Assessment is an important activity and needs to be done by teachers in schools. The conventional paradigm often interprets the assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syarifuddin, 2020). If referring to the current paradigm, assessment in schools is divided into three parts, namely assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). Assessment as learning has almost

the same function as assessment for learning but assessment as learning involves students in assessment, such as assessing themselves or colleagues. Assessment of learning is an assessment carried out after all learning ends and aims to assess student achievement. Assessment for learning is an assessment carried out during the learning process and aims to improve learning. It can play a role in preventing students from experiencing further learning failure because of its position between the other two assessments (Earl, 2013) as shown in the following assessment pyramid (Figure 1).



Figure 1. Assesment Pyramid

Assessment activities can be applied with tests. A test is a tool or procedure used to find out or measure students' abilities about something with certain rules (Arikunto, 2012). A test consists of two types, namely multiple choice and essay. Multiple choice test is a form of assessment in which each item provides an answer choice, and one of the choices is the correct answer. The essay test is a form of assessment that requires answers in sentences or words. Each type of test has strengths or weaknesses with each other. The strength of multiple-choice tests over essays is that multiple-choice tests can be conducted for many students, are more objective, and the test results can be known more quickly; but has a weakness, namely the multiple choice test is not able to see the actual abilities of students and the answers tend to guess or try it out (Rosidin, 2017). In addition, the strength of the multiple choice test has a scoring certainty compared to the essay test, namely 1 and 0 (Getting score of 1 for the correct answer, and score 0 for the wrong answer choice).

Multiple choice tests with only two answer choices are called dichotomous tests, and multiple choice tests with more than two answer choices are called polytomous tests (Kartono, 2008).

Until now, multiple choice tests are still widely used by teachers to assess students' abilities, especially students with a large number and wide area. To reduce the weakness of multiple choice tests, in the last four decades, experts have developed multiple choice tests by combining multiple choice tests and essays into multiple choice tests with reasons, and called polytomous tests with responses; or abbreviated the polytomous response test (Suwarto, 2012). The polytomous response test score is 1-4. Score 4 for the correct answer and reason, score 3 for correct answer but wrong reason, score 2 for wrong answer but correct reason, and score 1 for wrong answer and reason (Kartono, 2008).

In the 80s, the first time that was focused on and developed by experts was the closed polytomous response test (Treagust, 1988), and this test aims to diagnose misconceptions in biology, physics, and chemistry. This test consists of two levels; The first level is choosing answers on the multiple choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as tests on light and optical materials (Widiyatmoko & Shimizu, 2018), test on calculus material (Khiyarunnisa & Retnawati, 2018), test on acid and base material (Andaria & Hadiwinarto, 2020), test on reasoning material (Ambarwati et al., 2020), test on human digestive system material (Jamhari, 2021), test on mathematical connection material (Lestari et al., 2021). Meanwhile, the development of multiple-choice tests with open reasons is still limited, namely mathematics in calculus (Yang et al., 2017), and outside mathematics, namely physics in Higher Order Thinking Skills (Prasetya et al., 2019). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions

cannot be known in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), student answers are still guessing (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors is easy to observe (Treagust, 1988), and The suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test to be the open polytomous response test. The open polytomous response test is a form of multiple choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). Until now, research on the open polytomous response test is still very limited, such as tests on calculus material in universities (Yang et al., 2017), and tests on physics material in high school (Prasetya et al., 2019). Because there is still limited research on the open polytomous response test, it is necessary to conduct research on research subjects with other characteristics. This research was conducted on students in vocational schools who have different characteristics from students in college or high school. The characteristic difference is students in vocational schools place mathematics as a secondary subject, while students in colleges and high schools place mathematics as a primary subject (Oktaria, 2016). In addition, graduate students in vocational schools are more oriented towards practical abilities and skills, and in contrast to graduate students in colleges and high schools are academically oriented (Permendikbud, 2016). Therefore, it is reasonable to conduct research by developing the open polytomous response test for students in vocational schools. The test instrument developed must be accountable as a condition of a good test, and it is necessary to analyze the quality of the item (Rosidin, 2017).

There are two theories in analyzing item quality, namely classical and modern theory. Classical theory is a measurement theory for assessing tests based on the assumption of

measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). Modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely it cannot separate the characteristics of students and items, also the characteristics of items will change when students change. So, classical theory is considered less able to provide information about students' actual abilities. Modern theory is a solution to overcome the weaknesses of classical theory, because in modern theory an item does not affect other items (local independence), and items only measure one dimension/unidimensional (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable, the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

This research is a development research that aims to produce a good open polytomous response test according to classical and modern theory. The problem formulations proposed are (1) does the open polytomous response test have a good category according to classical and modern theory?, and (2) can the open polytomous response test provide information on the actual competence of students?.

Methodology

Research Design

This research is a research and development that refers to Plomp's model (Plomp, 2013) with the research procedure consisting of five stages, namely preliminary investigation, design,

realization or construction, test phase, revision, and implementation (test).

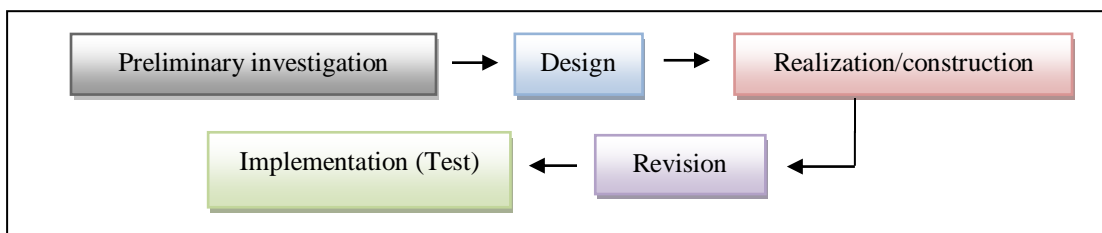


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the assessment instruments used by teachers so far. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make an expert assessment questionnaire sheet. The realization or construction stage is developing the items test, and also the expert validation process for the items test. The revision stage is the improvement of items test based on expert advice. The implementation (test) stage is to try out the test to students, and analyze the results of the test.

Research Subject

The subjects of the study are students of a vocational school in the province of Lampung, Indonesia. The research sample is determined using a non-probability sampling technique in the form of accidental sampling. It means taking the subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools are three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects are 413 students at grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 from the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
SMK Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collecting Technique

Data were collected using a questionnaire and test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed and to determine content validity (Suhaini et al., 2021). **The instrument was validated by two people** who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert are the suitability of the items with the indicators, language, and alternative answers to the questions. The score in validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining content validity, the instrument was tested on students. Then, it was continued by determining the validity of the construct and its reliability, with the aim that the instrument can be further analyzed.

The instrument used was the open polytomous response test which consisted of 40 questions

on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contains five answer choices along with the reasons. Student answers score refers to the polytomous score in the Partial Credit Model, where answer choices and reasons are related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Student Answers Score

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The research data obtained are analyzed in two stages, namely (1) questionnaire data analysis (qualitative analysis), and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis.

1. Questionnaire data analysis (qualitative analysis)

The questionnaire data analyzed include two parts, namely identification data on the instruments used by the teacher and expert validation data. The identification data on the instrument are analyzed descriptively, and the expert validation data are analyzed for trends or expert agreement using the Gregory formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range 0 - 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an

item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the value of the validity of an item.

2. *Test data analysis (empirical analysis)*

After the content validity test, it is followed by construct validity and reliability tests. The construct validation test uses exploratory factor analysis. The instrument is considered to have a good construct if the explained Kaiser-Meyer-Olkin (KMO) value is more than 0.5 (Retnawati, 2014). Reliability test using Cronbach's Alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further tests can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is because they are both preliminary analyzes of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the IteMan program is used for classical theory, and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program is used because it has several advantages (Untary et al., 2020), namely, it can analyze polytomous data and can analyze the maximum likelihood model using a 1-parameter logistic model.

2.1 *Analysis of test data with classical theory*

- a. The item difficulty level is the percentage of the number of students who answer the questions correctly or incorrectly. The difficulty of a good (medium) item is if the index is in the range of 0.3 to 0.7. If the index is below 0.3 then the item is difficult, and vice versa if the index is above 0.7 then the item is easy.
- b. Item discrimination is the item's ability to distinguish high-ability students from stupid, low-ability students. Good item discrimination if it has an index above 0.3; and if the item discrimination index is below 0.3 then the item needs to be revised

(Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items identifies the ability of about 50% of respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column.

According to modern theory, before analyzing the level of difficulty and item discrimination, three assumptions must be tested, namely unidimensional, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigen values of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals on the unidimensional solution (DeMars, 2010). In this study, the dimensional test uses the Eigen value analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by

other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer on each item. If the unidimensional assumption is accepted, the local independence assumption will automatically be accepted (DeMars, 2010). Model fit test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring Outfit Mean Square (MNSQ) and Pt-Measure. If the Outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it is found that so far the teacher has never used the polytomous response instrument with a multiple-choice test with open reasons. As many as 80% of teachers use essay tests and 20% of teachers use multiple choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers use this assessment as a learning improvement, such as improving lesson plans and teaching models/methods. The results of the questionnaire stated that 90% of teachers who do not use assessment as an improvement in learning are caused by several aspects, such as teachers are not understanding assessment (20%), teachers are not knowing how to analyze assessments (50%), and teachers are not knowing how to develop good assessment questions (30%). The following is a summary of the questionnaire data from the identification of assessment for learning instruments.

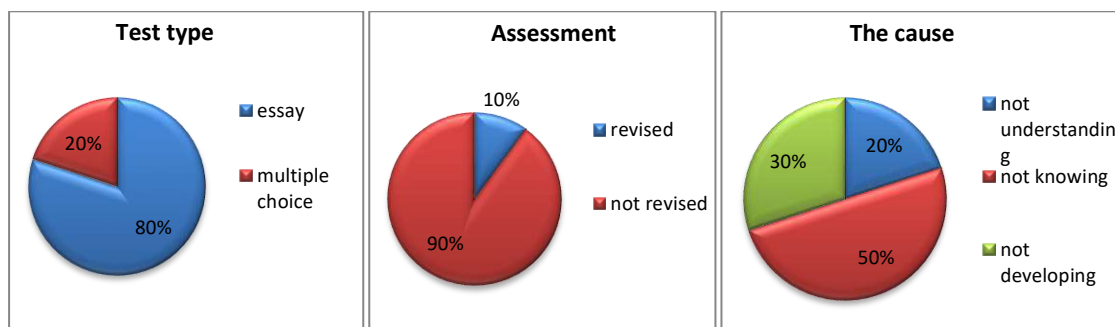


Figure 3. Description of Teacher Condition in Assessment for Learning

Content Validity

The results of the two expert assessments show that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement with the Gregory Index formula is obtained as shown in Table 4 below.

Table 4. Index Gregory Items

		Rater 1	
		weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provide some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arrange them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it is followed by a construct validity test. The results of the test

with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained the Cronbach's Alpha coefficient value of 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to the classical and modern methods.

Table 6. Item Reliability

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data with classical theory does not require testing assumptions, but the analysis of the level of difficulty and item discrimination can be directly calculated. The results of the analysis of the level of difficulty and item discrimination are obtained as shown in Table 7 below.

Table 7. Level of Difficulty and Item Discrimination

Item	Level of difficulty	Category	Discrimination	Category	Item	Level of difficulty	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised
4	0.540	good	0.304	Good	24	0.438	good	-0.071	revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	revised

7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	revised

Based on Table 7, it is found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are included in the good category. Meanwhile, only two items on discrimination have good categories, and the remaining items need to be revised. The results indicate that all items are good based on the level of difficulty, but almost all items need to be revised for item discrimination.

Analysis of Test Data with Modern Theory

Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, then calculating the eigenvalues. The eigenvalue is then used to calculate the percentage of explained variance, as well as describe the scree plot (Retnawati, 2014). The output of factor analysis is the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is based on the cumulative percentage of eigenvalues and scree plots. If the cumulative percentage of eigenvalues in the first factor is more than 20%, then the unidimensional assumption is accepted (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the eigenvalues in the first factor is 20.220%. The cumulative percentage of the eigenvalues has exceeded 20%, so the instrument in the study is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalues		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot which is based on the number of factors marked by the steepness of the graph with the acquisition of eigenvalues.

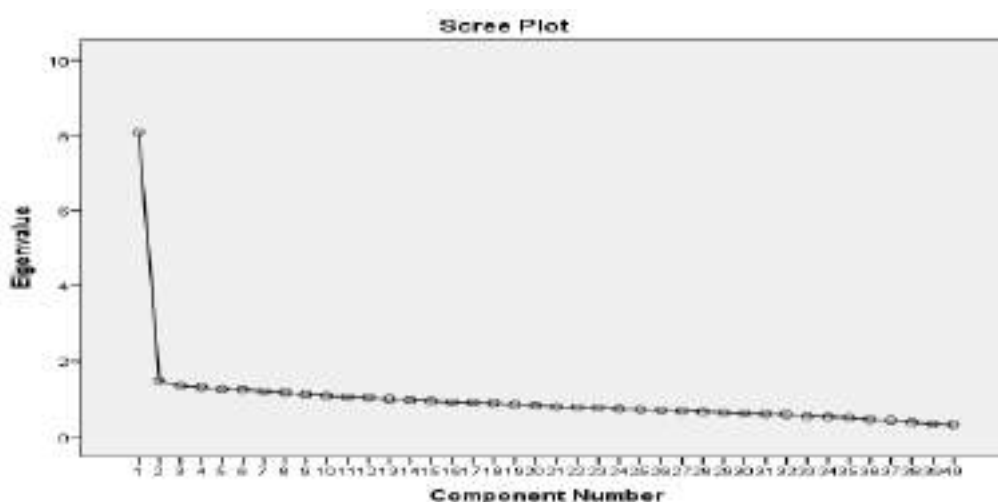


Figure 4. Scree Plot Unidimensi

Based on the scree plot, it is known that the eigenvalues immediately slope on the second factor. It shows that there is only one dominant factor in the developed instrument. The results prove that the test kit meets the unidimensional assumption or in other words only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be accepted if the respondent's answer to one item does not affect the respondent's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. It confirms that the assumption is automatically proven after being proven by the unidimensionality of the respondent's data on a test (Retnawati, 2014).

Table 10. Covariance Matrix

	<i>K1</i>	<i>K2</i>	<i>K3</i>	<i>K4</i>	<i>K5</i>	<i>K6</i>	<i>K7</i>	<i>K8</i>	<i>K9</i>	<i>K10</i>
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Data in Table 10 shows the results of the variance-covariance values between groups of students' abilities. It can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. It can be concluded that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test is analyzed using the Winsteps program. The item requirements are called fit to the model if the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Corr is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items match the model or fit (Table 11).

Table 11. Item Fit on Model

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73

Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-4	.98	-3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Difficulty Level

The item difficulty level is analyzed using the Winsteps program, and the results obtained can be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of -2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely the difficulty level is in the range of -2 and 2.

Table 12. Item Difficulty Level

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73

Person: REAL SEP.: 2.72 REL.: .58 ... Item: REAL SEP.: 3.66 REL.: .93

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBSN	MATCH EXPN	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-4	.98	-3	.75	.43	48.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	48.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-5	.97	-5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-4	.98	-4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	48.5	51.0	Q16
17	964	413	.07	.07	.96	-7	.96	-6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-4	.97	-4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-4	.97	-4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	48.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-5	.97	-5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.25	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.06	.07	1.04	.6	1.05	.8	.54	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-6	.96	-6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-3.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

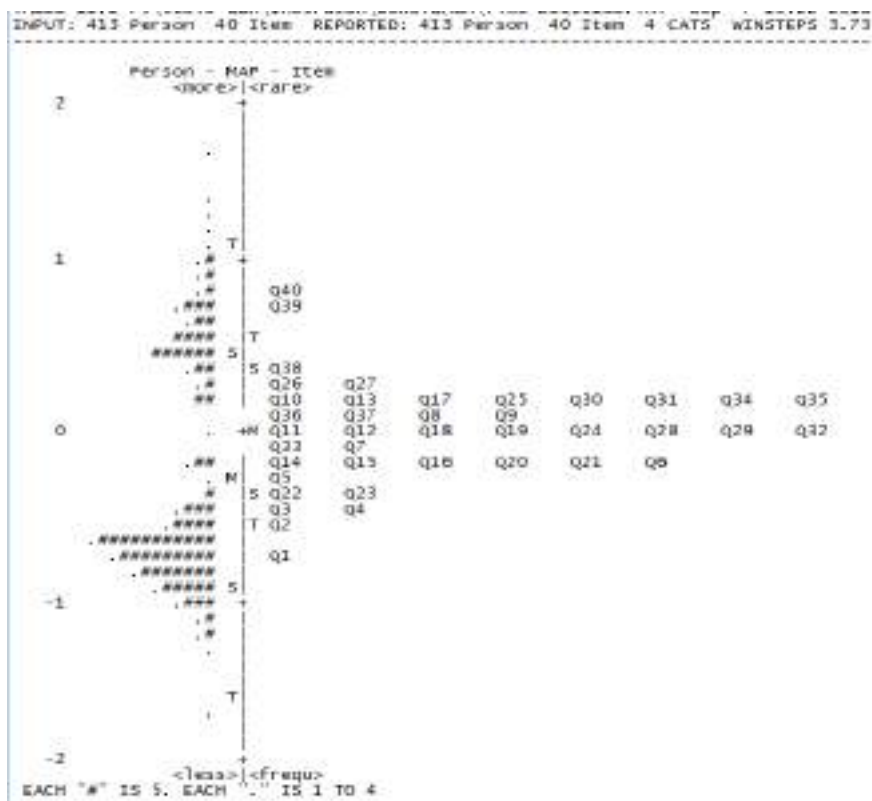


Figure 5. Item Difficulty Map

Item Discrimination

The item discrimination analysis used the Winsteps program and the results are presented in Table 13 (column PT-Measure Corr). Item discrimination is in the range of 0.23 to 0.74; with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

Table 13. Item Discrimination

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73

Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Comparative Analysis between Classical and Modern Theory

The results of the analysis of classical and modern theory obtained the index of difficulty level and item discrimination as follows.

Table 14. Analysis Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Category	Percentage	Many Items with Good Category	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 14, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has

more items in good category than using classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Table 13), it can be seen that there is a match between the categories of item discrimination. It means that if the item discrimination is not good with the classical theory, then the item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability is measured by using a test that is expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error or Standard Error Measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

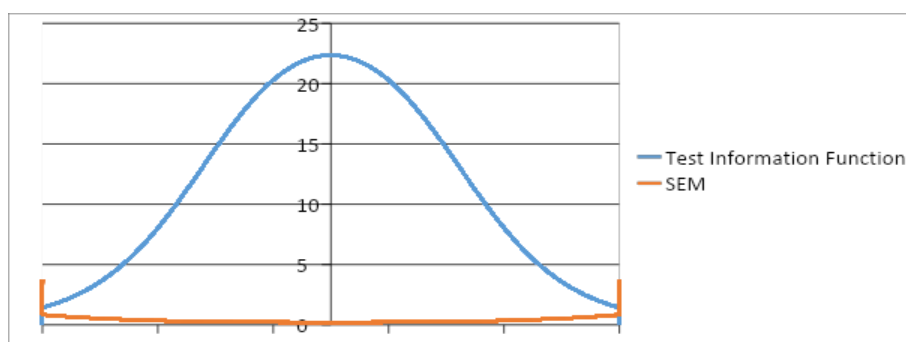


Figure 5. Graph of Information Function and Measurement Error

Figure 5 shows that the instrument provides a maximum of 22.36 information and has the smallest measurement error of 0.21 if it is given to students with moderate ability, which is 0.2. The lower limit and upper limit of the interval is the ability score where the graph of the information function and the SEM graph intersect in that interval. The graph indicates that

the greater the value of the information function, the smaller the measurement error (SEM). Item information function states the strength or contribution of test items in revealing the latent trait as measured by the test. With the item information function, it is known which items match the model, thus helping in the selection of test items (Retnawati, 2014). In conclusion, the characteristics of the test kit are suitable for students with moderate abilities.

Based on the test results, in addition to knowing the quality of the developed test instruments, it can also be seen the ability of students to work on the questions given. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of 6 student answers were selected as samples with different abilities (high, medium, and low).

Item 1: the cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms. arithmetic sequence.

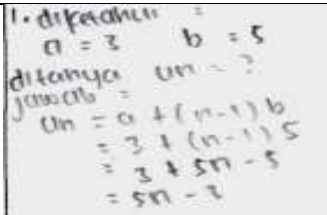
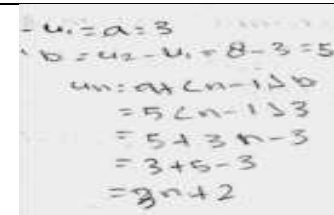
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	 <p>1. diketahui : $a = 3$ $b = 5$ ditanya : $U_n = ?$ Jawab : $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	 <p>$U_n = a + (n-1)b$ $b = u_2 - u_1 = 8 - 3 = 5$ $u_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>

Figure 6. Student Answers in Item 1

Item 2: the cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand and determine the number of terms in a sequence by using the general formula for an arithmetic

sequence or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

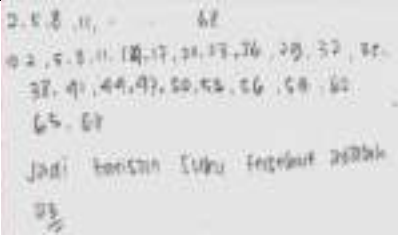
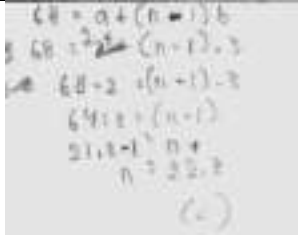
Question 2:	Pattern 1	Pattern 2
<p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>		

Figure 7. Student Answers in Item 2

Item 3: the cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and determined the difference or difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

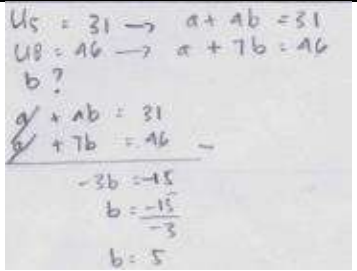
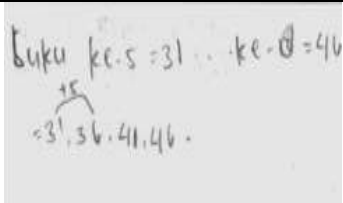
Question 3:	Pattern 1	Pattern 2
<p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>		

Figure 8. Student Answers in Item 3

Item 4: the cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and can determine

the n th term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula, but by writing the terms of the known terms and inserting several terms and continuing until the n th term.

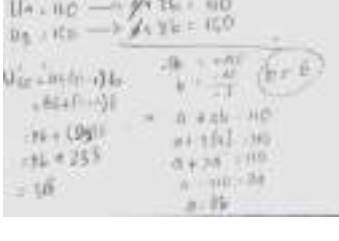
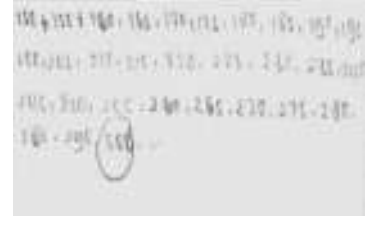
Question 4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>	 <p>Handwritten student solution for Pattern 1. The student uses the formula $U_n = a + (n-1)b$. They set up two equations: $U_4 = 110 \rightarrow a + 3b = 110$ and $U_9 = 150 \rightarrow a + 8b = 150$. They subtract the first equation from the second to get $5b = 40$, so $b = 8$. Then they substitute $b = 8$ into the first equation to get $a + 24 = 110$, so $a = 86$. Finally, they calculate $U_{30} = 86 + (30-1) \cdot 8 = 86 + 232 = 318$.</p>	 <p>Handwritten student solution for Pattern 2. The student lists terms of the sequence: $110, 118, 126, 134, 142, 150, 158, 166, 174, 182, 190, 198, 206, 214, 222, 230, 238, 246, 254, 262, 270, 278, 286, 294, 302, 310, 318$. They identify 318 as the 30th term.</p>

Figure 9. Student Answers in Item 4

Item 5: the cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but by writing the terms from known terms and inserts several terms and then defines them.

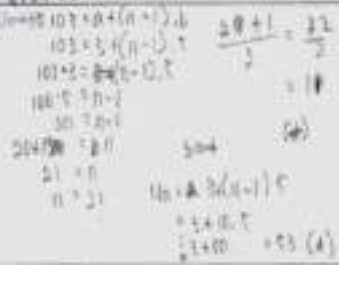

Question 5:	Answer Pattern 1	Answer Pattern 2
<p>An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...308</p> <p>A. 53 B. 52 C. 20 D. 11 E. 10</p> <p>Reason:</p>	 <p>Handwritten student solution for Answer Pattern 1. The student uses the formula $U_n = a + (n-1)b$. They set up the equation $103 = 3 + (n-1) \cdot 5$. They solve for n: $103 - 3 = 5(n-1) \rightarrow 100 = 5(n-1) \rightarrow 20 = n-1 \rightarrow n = 21$. They then calculate the middle term: $U_{21} = 3 + (21-1) \cdot 5 = 3 + 100 = 103$.</p>	 <p>Handwritten student solution for Answer Pattern 2. The student lists terms of the sequence: $3, 8, 13, 18, 23, 28, 33, 38, 43, 48, 53, 58, 63, 68, 73, 78, 83, 88, 93, 98, 103$. They identify 103 as the 21st term.</p>

Figure 10. Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test and all

parameters have been accepted. This instrument is a combination of multiple choice test and essay. Multiple choice tests are easier to check students' answers but students' mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find out deeper mathematical thinking processes but it takes a long time to check the answers.

Assessment of learning instruments that have been carried out with item analysis and the results of student analysis is one of the important sources of composite scores to be reported. In the final report, the test taker's ability score should be changed to a score of 0 - 10 from 0 - 100, according to the needs of the school. The transformation uses a linear transformation by dividing the score by the ideal score and then the result is multiplied by 10 to get a value in the range 0 – 10 or multiplied by 100 to get a score of 0 – 100. In the range 0 – 10, the score obtained by students taking the test the highest mathematics learning was 8.56 and the lowest was 4.31. In the range 0 – 100, the scores obtained by students with the highest mathematics is 85.625 and the lowest is 43.125.

The results of the assessment of students' mathematical abilities are presented in the form of very low to very high predicates. The results of the test analysis show that most students have low and very low abilities, namely 62% (253 students). Meanwhile, students who have high and very high abilities are 38% (160 students). The results of another analysis find that students who have high abilities tend to work according to the concepts that have been given by the teacher but do not follow the completion steps, students who have moderate abilities can solve problems according to the concepts that have been given by the teacher and the steps, and there are students who have abilities but are not able to use the concepts given by the teacher and are not even able to give clear reasons..

Another result of this study is that teachers agree to provide learning assessments with multiple choice questions with open-ended reasons because the instrument is easier for teachers to find out students' difficulties in certain materials. In this way, teachers can also provide remedial or other assistance to students who have learning difficulties. It means that the polytomous response instrument can be used as a way to determine which students need remedial or not. In general, previous research states how to determine students who need remedial only one test, namely multiple-choice tests (Gierl et al., 2017) or essays (Putri et al., 2020).

Discussion

This research is development research to produce the open polytomous response test. The instrument is a multiple choice test with open reasons. This instrument is analyzed by classical and modern theory. There are differences in the results of the analysis between classical and modern theories, namely item discrimination. Classical theory analysis obtains 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtains 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as not good but the results of the modern analysis are categorized as good, and vice versa (Retnawati, 2014). It means, if you find items that are not in a good category with classical theory, do not be in a hurry to revise or replace them before the analysis of modern theory analysis.

Research on assessment for learning with the open polytomous response test is still limited. When compared to previous research, there is only one study on assessment for learning with polytomous responses (Yang et al., 2017). However, Yang study has several fundamental differences, namely the research objectives and data analysis. The research aims to diagnose

student errors in university on the concept of calculus, not to produce a good assessment instrument. The data analysis uses parametric statistics (covariance), not using item analysis (classical and modern). Since the objectives and data analysis are different, the results of the study cannot be compared with the results of this study. However, this research has provided a reference for researchers in making reasoned multiple choice tests, such as the suitability of items with indicators, language, and alternative answers to questions.

Other studies are similar to assessment for learning with the open polytomous response test (Sarea, 2018). The similarity with Sarea's research lies in the research objectives and the analysis used (classic and modern). However, the difference is the researchers do not develop their questions and the questions are in the form of the closed polytomous response test. The results of Sarea's research states that the comparison of the results of the classical and modern theory of item analysis is different. The difference is that the level of difficulty and item discrimination in the classical theory is more categorized as good than the modern method. In other words, the modern way of stating the level of difficulty and item discrimination is categorized as good even though the analysis method states that the items are categorized as not good. Likewise with Saepuzaman's research the closed polytomous response test provide confidence that items that are not good according to classical theory are actually good items according to modern theory. The results of previous studies have provided support for the development of instruments on the polytomous response test, and this test instrument can be used as an alternative to all learning assessment (assessment: as learning, for learning, and of learning) for all vocational schools in Lampung, Indonesia and even outside Indonesia.

Conclusion

Based on the results of the research and discussion, conclusions are obtained, namely (1) the open polytomous response test have a good category according to classical and modern

theory. So, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. this is observed in the students' arguments in giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, teachers should familiarize students with giving a test in the form of a polytomous response before giving the test. For schools, principals or other leaders should encourage other teachers to take advantage of this test, and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important, so that students' prior knowledge can be known so that learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not been the researchers' expectations, for example representing schools with high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic material (sequences and series,

matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Ambarwati, R., Sunardi, Yudianto, E., Murtikusuma, R. P., & Safrida, L. N. (2020). Developing mathematical reasoning problems type two-tier multiple choice for junior high school students based on ethnomathematics of jember fashion carnival. In Suratno (Ed.) *ICOLSSSTEM*. IOP Publishing. <https://doi.org/10.1088/1742-6596/1563/1/012036>.
- Andaria, M. & Hadiwinarto. (2020). Development of a two-tier multiple choice question assessment instrument to measure students science process skills on acid-base material. *ISEJ: Indonesian Science Education Journal*, 1 (3), 257-268. <https://siducat.org/index.php/isej/article/view/141>.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi/ Journal of Mathematics, Statistics, & Computing*, 9(2), 95-113. <https://doi.org/10.20956/jmsk.v9i2.3402>.
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *Jurnal Suluh Pendidikan/Journal of Educational Suluh*, 17(1), 32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: the cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-Tier

- Diagnostic Test With Certainty of Response Index on The Concepts of Fluid. In C. Anwar (Ed.), *Young Scholar Symposium on Transdisciplinary in Education and Environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf>
- DeMars, C. E. (2010). *Item response theory*. Oxford University Press.
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Hong Kong: Assessment & Support Team. <https://wlts.edb.hkedcity.net/en/home/AandLI2.html>.
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87 (6), 1082–1116. <https://doi.org/10.3102/0034654317726529>.
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12 (3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543>.
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Jamhari, M. (2021). Developing the two-tier multiple choice tests in enhancing students' higher-order thinking skills on human digestive system. *Eduproxima: Jurnal Ilmiah Pendidikan IPA*, 3 (1), 50-64. <https://doi.org/10.29100/eduproxima.v3i1.1853>.
- Kartono. (2008). Equating the combined dichotomous and polytomous item test model in an achievement test. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>.
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how?. In A. W. Subianto (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <http://seminar.uny.ac.id/icriems/proceeding2018>.
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical

- connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://doi.org/10.31327/jme.v6i2.1607>.
- Linacre, J. M. (2012). *Winstep : Rasch-model computer programs*. Winsteps. Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset Pemasaran [Marketing Research]*. Eirlangga.
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. In Cari (Ed.), *Current research in Pandemic Covid-19 Era in Indonesia - 2nd ICOSETH 2020* (pp. 44-55). University of Sebelas Maret. <https://jurnal.uns.ac.id/ijsascs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools [Unpublished master's thesis]*. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah [Content standards for primary and secondary education]*. Indonesian Government publication service.
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>.
- Prasetya, A., Rosidin, U., & Herlina, K. (2019). Development of instrument assessment for learning the polytomous response models to train higher order thinking. In C. Anwar (Ed.), *Young Scholar Symposium on Transdisciplinary in Education and Environment (YSSTEE)* (pp. 1-11). UIN Raden Intan. <http://repository.lppm.unila.ac.id/11982/1/33.%20Prasetya%202019%20J.Phys.Conf.Ser..pdf>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://doi.org/10.15294/jere.v9i2.46133>.
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>.
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran [Learning evaluation and assessment]*. Media Akademi.
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisika dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst : Jurnal Pendidikan Fisika dan Terapannya/ Karst : Journal of Physics Education and Its Application*, 4 (1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>.
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam questions for Islamic

- religious education and character at the elementary school level]. *An-Nahdhah/An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article/view/40>.
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Classical tes theory vs response theory? *Didaktika: Jurnal Kependidikan/Didactics: Journal of Education*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>.
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic tests in learning]. Graha Ilmu.
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>.
- Treagust, D. (1988). Development and use of diagnostic tests to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169.
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep* [Analysis of research data with rash and winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122-1?inline=1>.
- Widiyatmoko, A., & Shimizu, K. (2018). The Development of two-tier multiple choice test to assess students' conceptual understanding about light and optical instruments. *Jurnal Pendidikan IPA Indonesia*, 7(4), 491-501. <https://doi.org/10.15294/jpii.v7i4.16591>.
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>.

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :

Class/Department :

School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

<p>1. Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ C. $U_n = 4n - 1$ B. $U_n = 5n - 2$ D. $U_n = 3n + 2$ C. $U_n = 4n - 1$</p> <p>Reason:</p>	<p>2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 D. 23 B. 13 E. 24 C. 22</p> <p>Reason:</p>
<p>3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 D. 8 B. 6 E. 11 C. 7</p> <p>Reason:</p>	<p>4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 D. 344 B. 318 E. 354 C. 326</p> <p>Reason:</p>
<p>5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...</p> <p>A. 53 D. 11 B. 52 E. 10 C. 20 D. 11 E. 10</p> <p>Reason:</p>	<p>6. Given the arithmetic sequence: 4, 10, 16, 22, If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...</p> <p>A. 18 D. 24 B. 20 E. 26 C. 22</p> <p>Reason:</p>
<p>7. The nth term of an arithmetic series is $U_n = 3n - 5$. The formula for the sum of the first n terms of the series is ...</p> <p>A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$ B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$ C. $S_n = \frac{n}{2}(3n - 4)$</p> <p>Reason:</p>	<p>8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....</p> <p>A. $5n - 20$ D. $2n - 20$ B. $5n - 10$ E. $2n - 10$ C. $2n - 30$</p> <p>Reason:</p>
<p>9. The sum of all integers between 100 and 300 which are divisible by 5 is ...$S_n = \frac{n}{2}(3n - 7)$</p> <p>A. 8.200 D. 7.600 B. 8.000 E. 7.400 C. 7.800</p> <p>Reason:</p>	<p>10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?</p> <p>A. 24 D. 27 B. 25 E. 28 C. 26</p> <p>Reason:</p>
<p>11. The middle term of an arithmetic sequence is 25. If the difference is 4 and the 5th term is 21. Then the sum of all the terms in the sequence is ...</p> <p>A. 175 D. 295 B. 189 E. 375 C. 275</p> <p>Reason:</p>	<p>12. A number of candies are distributed among five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces</p> <p>A. 60 D. 75 B. 65 E. 80 C. 70</p> <p>Reason:</p>
<p>13. The sum of the first n terms of a series is $2n^2 - n$.</p>	<p>14. The number of terms in the geometric sequence:</p>

Thanks, received.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

[Quoted text hidden]



SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>

3rd round corrections request for the manuscript ID# 21112502244011

3 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>
Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Wed, Mar 16, 2022 at 7:00 PM

Dear Dr. Sutiarso,

Please see the attached file as the second round corrections.

Please remove the old highlights and re-highlight for new edited parts. We need a new correction report.

We are looking forward to getting your second revised paper until **March 30, 2022**.**PS. If the all corrections can't be done, the editorial process will be cancelled.**

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com

On 15-Mar-22 3:14 PM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research.

I apologize for my error in citation (there is a problem in my computer). Here I re-send my article.
Thank you for this opportunity to improve.

Best regards,
Sugeng Sutiarso
Lampung University, Indonesia.

On Tue, Mar 15, 2022 at 1:57 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

In-text citations are not visible in the edited file (we guess it's because of the program you were using). Could you correct the citations and re-send it please urgently?

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 13-Mar-22 3:51 PM, SUGENG SUTIARSO wrote:

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.

I have revised the second round of corrections according to the reviewer's suggestion.
Here I attach a correction report and revised article.

Thank you.

Best regards,

Sugeng Sutiarso
Lampung University

On Mon, Feb 28, 2022 at 7:59 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarmo,

Please see the attached file as the second round corrections.

Please remove the old highlights and re-highlight for new edited parts. We need a new correction report.

We are looking forward to getting your second revised paper until **March 14, 2022**.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com

On 22-Feb-22 4:48 PM, SUGENG SUTIARSO wrote:


Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.


I have revised the article according to the reviewer's suggestion. Here I attach (1) a revised article, (2) a correction report, and (3) a proofreading certificate from my university's language center.

Best regards,

Sugeng Sutiarmo
Lampung University

2 attachments

 **3rd round_EU-JER_21112502244011_R2612.doc**
170K

 **3rd round_EU-JER_21112502244011_R2613.docx**
1352K

SUGENG SUTIARSO <sugeng.sutiarmo@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Tue, Mar 29, 2022 at 4:25 PM

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.

I have revised the article according to the reviewer's suggestion. Here I attach (1) 3rd round of article revisions, and (2) 3rd correction report.

Best regards,

Sugeng Sutiarmo
Lampung University

[Quoted text hidden]

2 attachments

 **CORRECTION REPORT 3_Article Sugeng Sutiarmo et al.docx**
241K

 **3rd Revision_Article Sugeng Sutiarmo et al.docx**
1349K

Editor - European Journal of Educational Research <editor@eu-jer.com>

Thu, Mar 31, 2022 at 3:04 PM



European Journal of Educational Research

ISSN: 2165-8714

<http://www.eu-jer.com/>

Review Form

Manuscript ID:	2nd Revision_EU-JER_21112502244011	Date: 16 March 2022			
Manuscript Title:	The Development of an Assessment Instrument Using Polytomous Response in Mathematics				
ABOUT MANUSCRIPT (Mark with "X" one of the options)		Accept	Weak	Refuse	Not Available
Language is clear and correct			X		
Literature is well written			X		
References are cited as directed by APA		X			
The research topic is significant to the field			X		
The article is complete, well organized and clearly written			X		
Research design and method is appropriate		X			
Analyses are appropriate to the research question		X			
Results are clearly presented			X		
A reasonable discussion of the results is presented			X		
Conclusions are clearly stated			X		
Recommendations are clearly stated			X		
GENERAL REMARKS AND RECOMMENDATIONS TO THE AUTHOR					
<p>The flaws that I emphasized in my previous reviews are still continuing. There are:</p> <ol style="list-style-type: none"> 1- The rationality of this study is not still clear and explained based on the literature 2- The introduction is still long. 3- The statements in the introduction do not serve to explain the research problem. 4- Although the authors did write "Because there is still limited research on the open polytomous response test, it is necessary to conduct research..", the strengths and weaknesses of prior studies have not been emphasized in the introduction. 5- What I as to the authors, what are differences between the polytomous and two-tier tests? 6- In the discussion, remove these words "...do not be in a hurry to revise.." 7- The discussion has too few references to discuss the results. Also, it is too limited. 8- The problems with the use of the English language are continuing. This paper needs copyediting by a native speaker and an expert in the field. 					
THE DECISION (Mark with "X" one of the options)					
Accepted: Correction not required					
Accepted: Minor correction required					
Conditionally Accepted: Major Correction Required (Need second review after corrections)					X



European Journal of Educational Research

ISSN: 2165-8714

<http://www.eu-jer.com/>

Refused

Reviewer Code: R2612 (The name of referee is hidden because of blind review)

The Development of an Assessment Instrument Using Polytomous Response in Mathematics

Abstract: This research is development research aimed to produce a good instrument of assessment in mathematics using polytomous response according to classical and modern theories. This research design uses the Plomp model which consists of five stages, namely: preliminary investigation, design, realization or construction, trial, revision, and implementation (testing). The study was conducted in three vocational schools in Lampung, Indonesia. The study involves 413 students consisting of 191 male students and 222 female students. The data were collected through questionnaires and tests. Questionnaire was to identify instruments commonly used by teachers so far and to validate instruments by experts. The test used multiple-choice tests with open reasons as many as 40 items. The data were analyzed in two ways, namely analysis with classical and modern theories. The results show that the open polytomous response test have a good category according to classical and modern theory, and the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: assessment for learning, classical and modern theory, multiple choice tests with open reason, polytomous response, vocational school

Introduction

Assessment is an important activity and needs to be done by teachers in schools. The conventional paradigm often interprets the assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). If referring to the current paradigm, assessment in schools is divided into three parts, namely assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). Assessment as learning has almost the

same function as assessment for learning but assessment as learning involves students in assessment, such as assessing themselves or colleagues. Assessment of learning is an assessment carried out after all learning ends and aims to assess student achievement. Assessment for learning is an assessment carried out during the learning process and aims to improve learning. It can play a role in preventing students from experiencing further learning failure because of its position between the other two assessments (Earl, 2013) as shown in the following assessment pyramid (Figure 1).



Figure 1. Assessement Pyramid

Assessment activities can be applied with tests. A test is a tool or procedure used to find out or measure students' abilities about something with certain rules (Arikunto, 2012). A test consists of two types, namely multiple choice and essay. Multiple choice test is a form of assessment in which each item provides an answer choice, and one of the choices is the correct answer. The essay test is a form of assessment that requires answers in sentences or words. Each type of test has strengths or weaknesses with each other. The strength of multiple-choice tests over essays is that multiple-choice tests can be conducted for many students, are more objective, and the test results can be known more quickly; but has a weakness, namely the multiple choice test is not able to see the actual abilities of students and the answers tend to guess or try it out (Rosidin, 2017). In addition, the strength of the multiple choice test has a scoring certainty compared to the essay test, namely 1 and 0 (Getting score of 1 for the correct answer, and score 0 for the wrong answer choice). Multiple choice tests

with only two answer choices are called dichotomous tests, and multiple choice tests with more than two answer choices are called polytomous tests(Kartono, 2008).

Until now, multiple choice tests are still widely used by teachers to assess students' abilities, especially students with a large number and wide area. To reduce the weakness of multiple choice tests, in the last four decades, experts have developed multiple choice tests by combining multiple choice tests and essays into multiple choice tests with reasons, and called polytomous tests with responses; or abbreviated the polytomous response test(Suwarto, 2012). The polytomous response test score is 1-4. Score 4 for the correct answer and reason, score 3 for correct answer but wrong reason, score 2 for wrong answer but correct reason, and score 1 for wrong answer and reason(Kartono, 2008).

In the 80s, the first time that was focused on and developed by experts was the closed polytomous response test(Treagust, 1988), and this test aims to diagnose misconceptions in biology, physics, and chemistry. This test consists of two levels; The first level is choosing answers on the multiple choice test, and the second level is choosing reasons based on the answer choices at the first level(Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as tests on light and optical materials (Widiyatmoko & Shimizu, 2018), test on calculus material(Khiyarunnisa & Retnawati, 2018), test on acid and base material(Andaria & Hadiwinarto, 2020), test on reasoning material(Ambarwati et al., 2020), test on human digestive system material(Jamhari, 2021), test on mathematical connection material(Lestari et al., 2021). Meanwhile, the development of multiple-choice tests with open reasons is still limited, namely mathematics in calculus(Yang et al., 2017), and outside mathematics, namely physics in Higher Order Thinking Skills(Prasetya et al., 2019). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students'

misconceptions cannot be known in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), student answers are still guessing (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors is easy to observe (Treagust, 1988), and The suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test to be the open polytomous response test. The open polytomous response test is a form of multiple choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). Until now, research on the open polytomous response test is still very limited, such as tests on calculus material in universities (Yang et al., 2017), and tests on physics material in high school (Prasetya et al., 2019). Because there is still limited research on the open polytomous response test, it is necessary to conduct research on research subjects with other characteristics. This research was conducted on students in vocational schools who have different characteristics from students in college or high school. The characteristic difference is students in vocational schools place mathematics as a secondary subject, while students in colleges and high schools place mathematics as a primary subject (Oktaria, 2016). In addition, graduate students in vocational schools are more oriented towards practical abilities and skills, and in contrast to graduate students in colleges and high schools are academically oriented (Permendikbud, 2016). Therefore, it is reasonable to conduct research by developing the open polytomous response test for students in vocational schools. The test instrument developed must be accountable as a condition of a good test, and it is necessary to analyze the quality of the item (Rosidin, 2017).

There are two theories in analyzing item quality, namely classical and modern theory. Classical theory is a measurement theory for assessing tests based on the assumption of

measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). Modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely it cannot separate the characteristics of students and items, also the characteristics of items will change when students change. So, classical theory is considered less able to provide information about students' actual abilities. Modern theory is a solution to overcome the weaknesses of classical theory, because in modern theory an item does not affect other items (local independence), and items only measure one dimension/unidimensional (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable, the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

This research is a development research that aims to produce a good open polytomous response test according to classical and modern theory. The problem formulations proposed are (1) does the open polytomous response test have a good category according to classical and modern theory?, and (2) can the open polytomous response test provide information on the actual competence of students?.

Methodology

Research Design

This research is a research and development that refers to Plomp's model (Plomp, 2013) with the research procedure consisting of five stages, namely preliminary investigation, design,

realization or construction, test phase, revision, and implementation (test).

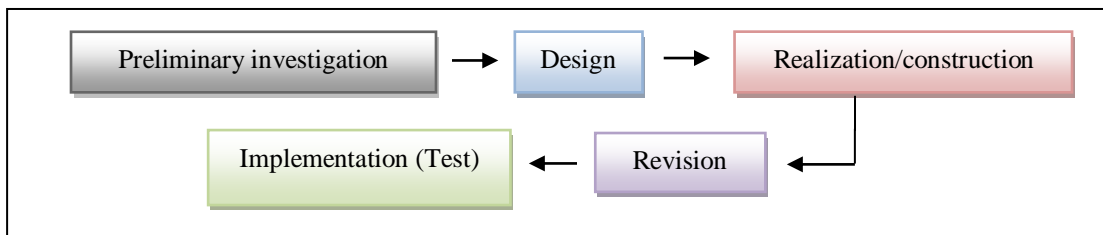


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the assessment instruments used by teachers so far. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make an expert assessment questionnaire sheet. The realization or construction stage is developing the items test, and also the expert validation process for the items test. The revision stage is the improvement of items test based on expert advice. The implementation (test) stage is to try out the test to students, and analyze the results of the test.

Research Subject

The subjects of the study are students of a vocational school in the province of Lampung, Indonesia. The research sample is determined using a non-probability sampling technique in the form of accidental sampling. It means taking the subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools are three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects are 413 students at grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 from the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
SMK Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collecting Technique

Data were collected using a questionnaire and test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed and to determine content validity(Suhaini et al., 2021). **The instrument was validated by two people** who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert are the suitability of the items with the indicators, language, and alternative answers to the questions. The score in validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining content validity, the instrument was tested on students. Then, it was continued by determining the validity of the construct and its reliability, with the aim that the instrument can be further analyzed.

The instrument used was the open polytomous response test which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and

functions, and matrices. Each item contains five answer choices along with the reasons. Student answers score refers to the polytomous score in the Partial Credit Model, where answer choices and reasons are related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Student Answers Score

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The research data obtained are analyzed in two stages, namely (1) questionnaire data analysis (qualitative analysis), and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis.

1. Questionnaire data analysis (qualitative analysis)

The questionnaire data analyzed include two parts, namely identification data on the instruments used by the teacher and expert validation data. The identification data on the instrument are analyzed descriptively, and the expert validation data are analyzed for trends or expert agreement using the Gregory formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range 0 - 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the value of

the validity of an item.

2. *Test data analysis (empirical analysis)*

After the content validity test, it is followed by construct validity and reliability tests. The construct validation test uses exploratory factor analysis. The instrument is considered to have a good construct if the explained Kaiser-Meyer-Olkin (KMO) value is more than 0.5 (Retnawati, 2014). Reliability test using Cronbach's Alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further tests can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is because they are both preliminary analyzes of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the IteMan program is used for classical theory, and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program is used because it has several advantages (Untary et al., 2020), namely, it can analyze polytomous data and can analyze the maximum likelihood model using a 1-parameter logistic model.

2.1 *Analysis of test data with classical theory*

- a. The item difficulty level is the percentage of the number of students who answer the questions correctly or incorrectly. The difficulty of a good (medium) item is if the index is in the range of 0.3 to 0.7. If the index is below 0.3 then the item is difficult, and vice versa if the index is above 0.7 then the item is easy.
- b. Item discrimination is the item's ability to distinguish high-ability students from stupid, low-ability students. Good item discrimination if it has an index above 0.3; and if the item discrimination index is below 0.3 then the item needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items identifies the ability of about 50% of respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column.

According to modern theory, before analyzing the level of difficulty and item discrimination, three assumptions must be tested, namely unidimensional, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigen values of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals on the unidimensional solution (DeMars, 2010). In this study, the dimensional test uses the Eigen value analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer on each item. If the

unidimensional assumption is accepted, the local independence assumption will automatically be accepted (DeMars, 2010). Model fit test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring Outfit Mean Square (MNSQ) and Pt-Measure. If the Outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it is found that so far the teacher has never used the polytomous response instrument with a multiple-choice test with open reasons. As many as 80% of teachers use essay tests and 20% of teachers use multiple choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers use this assessment as a learning improvement, such as improving lesson plans and teaching models/methods. The results of the questionnaire stated that 90% of teachers who do not use assessment as an improvement in learning are caused by several aspects, such as teachers are not understanding assessment (20%), teachers are not knowing how to analyze assessments (50%), and teachers are not knowing how to develop good assessment questions (30%). The following is a summary of the questionnaire data from the identification of assessment for learning instruments.

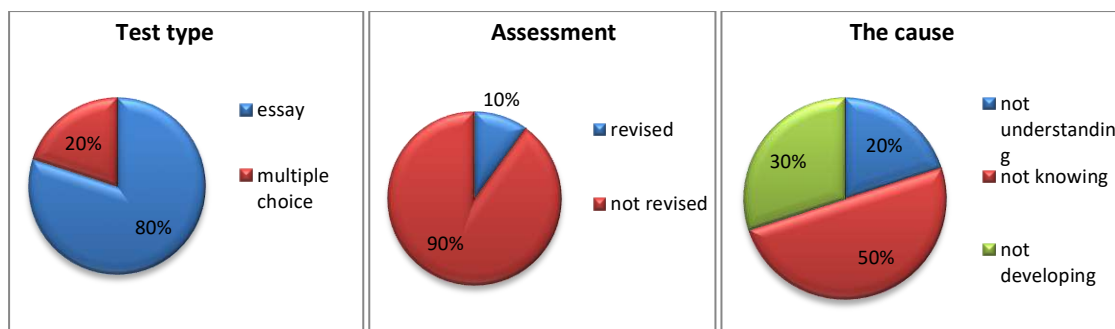


Figure 3. Description of Teacher Condition in Assessment for Learning

Content Validity

The results of the two expert assessments show that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement with the Gregory Index formula is obtained as shown in Table 4 below.

Table 4. Index Gregory Items

		Rater 1	
		weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provide some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arrange them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it is followed by a construct validity test. The results of the test

with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained the Cronbach's Alpha coefficient value of 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to the classical and modern methods.

Table 6. Item Reliability

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data with classical theory does not require testing assumptions, but the analysis of the level of difficulty and item discrimination can be directly calculated. The results of the analysis of the level of difficulty and item discrimination are obtained as shown in Table 7 below.

Table 7. Level of Difficulty and Item Discrimination

Item	Level of difficulty	Category	Discrimination	Category	Item	Level of difficulty	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised
4	0.540	good	0.304	Good	24	0.438	good	-0.071	revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	revised

8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	revised

Based on Table 7, it is found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are included in the good category. Meanwhile, only two items on discrimination have good categories, and the remaining items need to be revised. The results indicate that all items are good based on the level of difficulty, but almost all items need to be revised for item discrimination.

Analysis of Test Data with Modern Theory

Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, then calculating the eigenvalues. The eigenvalue is then used to calculate the percentage of explained variance, as well as describe the scree plot (Retnawati, 2014). The output of factor analysis is the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is based on the cumulative percentage of eigenvalues and scree plots.

If the cumulative percentage of eigenvalues in the first factor is more than 20%, then the unidimensional assumption is accepted (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the eigenvalues in the first factor is 20.220%. The cumulative percentage of the eigenvalues has exceeded 20%, so the instrument in the study is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalues		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot which is based on the number of factors marked by the steepness of the graph with the acquisition of eigenvalues.

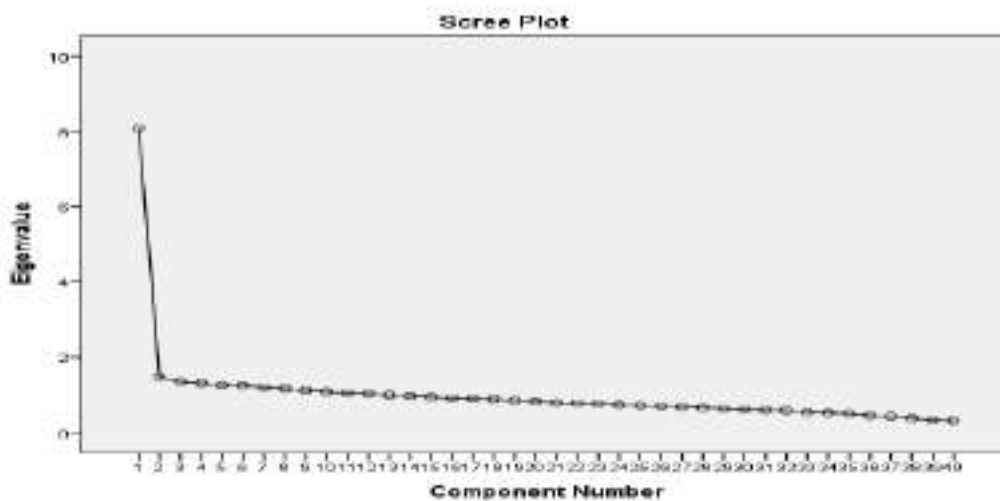


Figure 4. Scree Plot Unidimensi

Based on the scree plot, it is known that the eigenvalues immediately slope on the second factor. It shows that there is only one dominant factor in the developed instrument. The results prove that the test kit meets the unidimensional assumption or in other words only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be accepted if the respondent's answer to one

item does not affect the respondent's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. It confirms that the assumption is automatically proven after being proven by the unidimensionality of the respondent's data on a test(Retnawati, 2014).

Table 10. Covariance Matrix

	<i>K1</i>	<i>K2</i>	<i>K3</i>	<i>K4</i>	<i>K5</i>	<i>K6</i>	<i>K7</i>	<i>K8</i>	<i>K9</i>	<i>K10</i>
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Data in Table 10 shows the results of the variance-covariance values between groups of students' abilities. It can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. It can be concluded that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test is analyzed using the Winsteps program. The item requirements are called fit to the model if the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Corr is positive, then it can be said that the item fits the model(Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items match the model or fit (Table 11).

Table 11. Item Fit on Model

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73

Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93

Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Difficulty Level

The item difficulty level is analyzed using the Winsteps program, and the results obtained can be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of -2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely the difficulty level is in the range of -2 and 2.

Table 12. Item Difficulty Level

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73
 Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 5.66 REL.: .03
 Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBSN	MATCH EXPN	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q5
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.0	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.10	2.4	1.10	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	48.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.36	.44	41.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.14	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	48.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.25	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	953	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.60	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

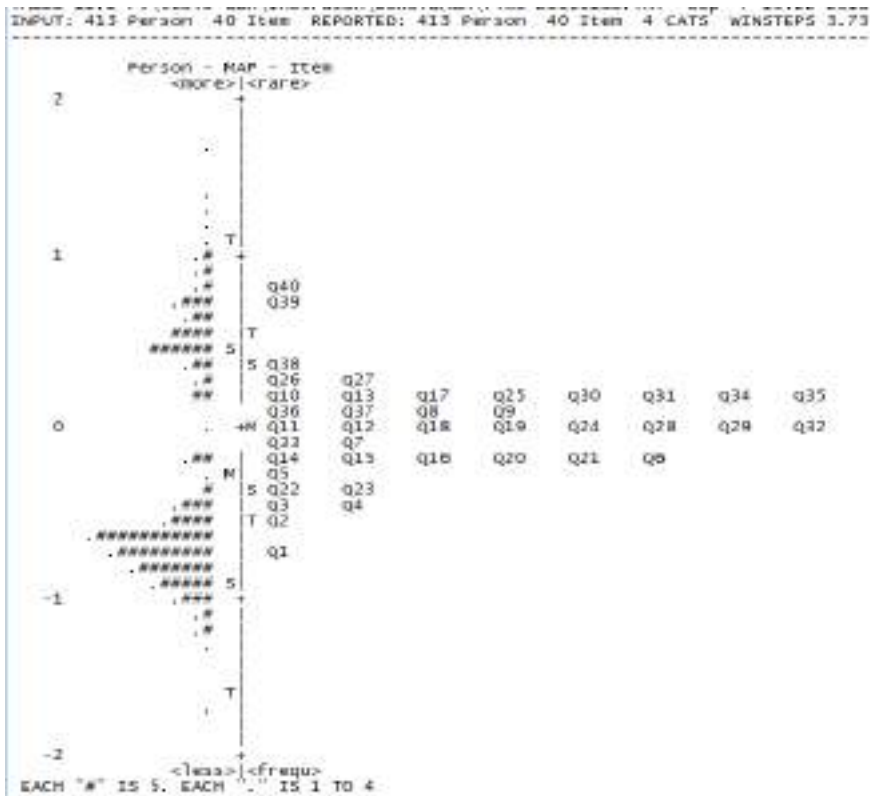


Figure 5. Item Difficulty Map

Item Discrimination

The item discrimination analysis used the Winsteps program and the results are presented in Table 13 (column PT-Measure Corr). Item discrimination is in the range of 0.23 to 0.74; with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

Table 13. Item Discrimination

INPUT: 413 Person 40 Item REPORTED: 413 Person 40 Item 4 CATS WINSTEPS 3.73
 Person: REAL SEP.: 2.72 REL.: .88 ... Item: REAL SEP.: 3.66 REL.: .93
 Item STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Comparative Analysis between Classical and Modern Theory

The results of the analysis of classical and modern theory obtained the index of difficulty level and item discrimination as follows.

Table 14. Analysis Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Category	Percentage	Many Items with Good Category	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 14, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in good category than using classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Table 13), it can be seen that there is a match between the categories of item discrimination. It means that if the item discrimination is not good with the classical theory, then the item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability is measured by using a test that is expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error or Standard Error Measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

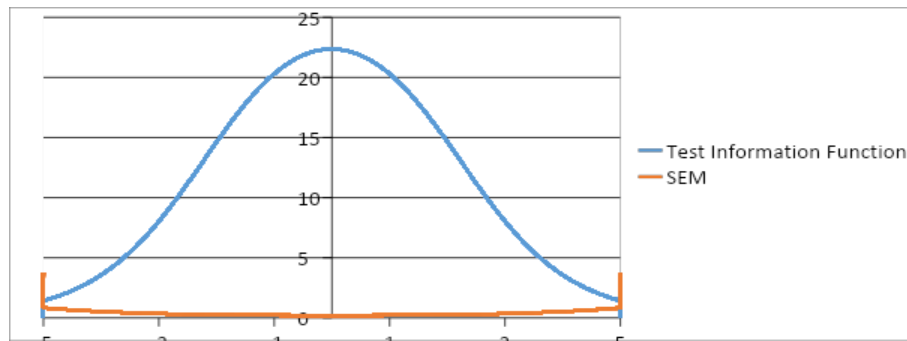


Figure 5. Graph of Information Function and Measurement Error

Figure 5 shows that the instrument provides a maximum of 22.36 information and has the smallest measurement error of 0.21 if it is given to students with moderate ability, which is 0.2. The lower limit and upper limit of the interval is the ability score where the graph of the information function and the SEM graph intersect in that interval. The graph indicates that the greater the value of the information function, the smaller the measurement error (SEM). Item information function states the strength or contribution of test items in revealing the latent trait as measured by the test. With the item information function, it is known which items match the model, thus helping in the selection of test items (Retnawati, 2014). In conclusion, the characteristics of the test kit are suitable for students with moderate abilities.

Based on the test results, in addition to knowing the quality of the developed test instruments, it can also be seen the ability of students to work on the questions given. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of 6 student answers were selected as samples with different abilities (high, medium, and low).

Item 1: the cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform

algebraic operations on general forms. arithmetic sequence.

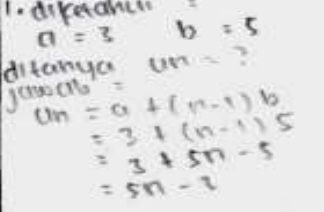
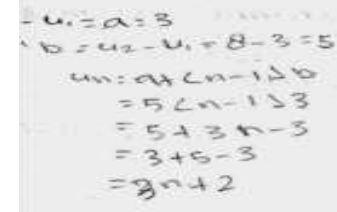
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>		

Figure 6. Student Answers in Item 1

Item 2: the cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand and determine the number of terms in a sequence by using the general formula for an arithmetic sequence or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

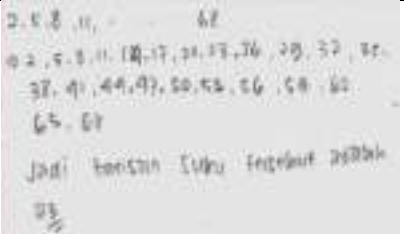
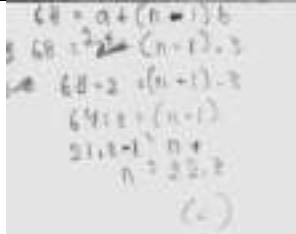
Question2:	Pattern 1	Pattern 2
<p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>		

Figure 7. Student Answers in Item 2

Item 3: the cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and determined the difference or difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the

terms of the known terms and inserting several terms.

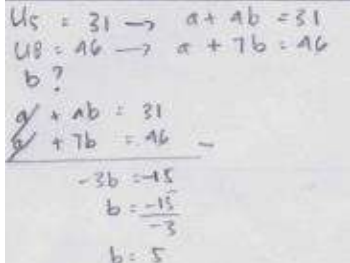
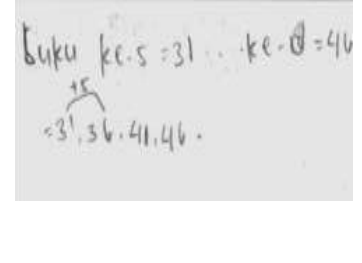
Question 3:	Pattern 1	Pattern 2
<p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>		

Figure 8. Student Answers in Item 3

Item 4: the cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and can determine the n th term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula, but by writing the terms of the known terms and inserting several terms and continuing until the n th term.

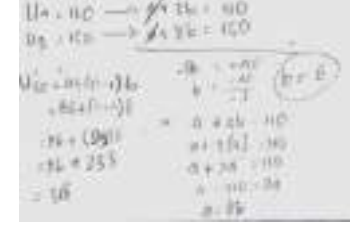
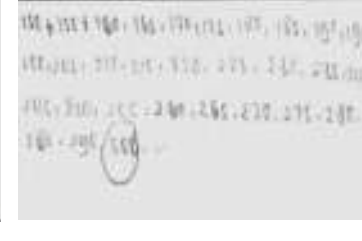
Question 4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>		

Figure 9. Student Answers in Item 4

Item 5: the cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but by writing the terms from known terms and inserts several terms and then defines them.

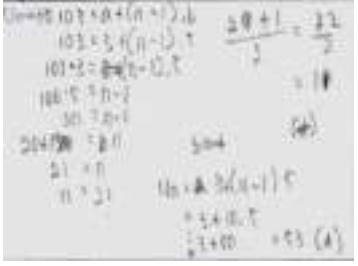

Question 5:	Answer Pattern 1	Answer Pattern 2
<p>An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...308</p> <p>A. 53 B. 52 C. 20 D. 11 E. 10</p> <p>Reason:</p>		

Figure 10. Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test and all parameters have been accepted. This instrument is a combination of multiple-choice test and essay. Multiple-choice tests are easier to check students' answers but students' mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find out deeper mathematical thinking processes but it takes a long time to check the answers.

Assessment of learning instruments that have been carried out with item analysis and the results of student analysis is one of the important sources of composite scores to be reported. In the final report, the test taker's ability score should be changed to a score of 0 - 10 from 0 - 100, according to the needs of the school. The transformation uses a linear transformation by dividing the score by the ideal score and then the result is multiplied by 10 to get a value in the range 0 - 10 or multiplied by 100 to get a score of 0 - 100. In the range 0 - 10, the score obtained by students taking the test the highest mathematics learning was 8.56 and the lowest was 4.31. In the range 0 - 100, the scores obtained by students with the highest mathematics is 85.625 and the lowest is 43.125.

The results of the assessment of students' mathematical abilities are presented in the form of very low to very high predicates. The results of the test analysis show that most students

have low and very low abilities, namely 62% (253 students). Meanwhile, students who have high and very high abilities are 38% (160 students). The results of another analysis find that students who have high abilities tend to work according to the concepts that have been given by the teacher but do not follow the completion steps, students who have moderate abilities can solve problems according to the concepts that have been given by the teacher and the steps, and there are students who have abilities but are not able to use the concepts given by the teacher and are not even able to give clear reasons..

Another result of this study is that teachers agree to provide learning assessments with multiple choice questions with open-ended reasons because the instrument is easier for teachers to find out students' difficulties in certain materials. In this way, teachers can also provide remedial or other assistance to students who have learning difficulties. It means that the polytomous response instrument can be used as a way to determine which students need remedial or not. In general, previous research states how to determine students who need remedial only one test, namely multiple-choice tests (Gierl et al., 2017) or essays (Putri et al., 2020).

Discussion

This research is development research to produce the open polytomous response test. The instrument is a multiple choice test with open reasons. This instrument is analyzed by classical and modern theory. There are differences in the results of the analysis between classical and modern theories, namely item discrimination. Classical theory analysis obtains 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtains 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as not good but the results of the modern analysis are categorized as good, and vice versa (Retnawati, 2014). It means, if you

find items that are not in a good category with classical theory, do not be in a hurry to revise or replace them before the analysis of modern theory analysis.

Research on assessment for learning with the open polytomous response test is still limited. When compared to previous research, there is only one study on assessment for learning with polytomous responses (Yang et al., 2017). However, Yang study has several fundamental differences, namely the research objectives and data analysis. The research aims to diagnose student errors in university on the concept of calculus, not to produce a good assessment instrument. The data analysis uses parametric statistics (covariance), not using item analysis (classical and modern). Since the objectives and data analysis are different, the results of the study cannot be compared with the results of this study. However, this research has provided a reference for researchers in making reasoned multiple choice tests, such as the suitability of items with indicators, language, and alternative answers to questions.

Other studies are similar to assessment for learning with the open polytomous response test (Sarea, 2018). The similarity with Sarea's research lies in the research objectives and the analysis used (classic and modern). However, the difference is the researchers do not develop their questions and the questions are in the form of the closed polytomous response test. The results of Sarea's research states that the comparison of the results of the classical and modern theory of item analysis is different. The difference is that the level of difficulty and item discrimination in the classical theory is more categorized as good than the modern method. In other words, the modern way of stating the level of difficulty and item discrimination is categorized as good even though the analysis method states that the items are categorized as not good. Likewise with Saepuzaman's research the closed polytomous response test provide confidence that items that are not good according to classical theory are actually good items according to modern theory. The results of previous studies have provided support for the

development of instruments on the polytomus response test, and this test instrument can be used as an alternative to all learning assessment (assessment: as learning, for learning, and of learning) for all vocational schools in Lampung, Indonesia and even outside Indonesia.

Conclusion

Based on the results of the research and discussion, conclusions are obtained, namely (1) the open polytomous response test have a good category according to classical and modern theory. So, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. this is observed in the students' arguments in giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, teachers should familiarize students with giving a test in the form of a polytomous response before giving the test. For schools, principals or other leaders should encourage other teachers to take advantage of this test, and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important, so that students' prior knowledge can be known so that learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not been the researchers' expectations, for example representing schools with high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic material (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Ambarwati, R., Sunardi, Yudianto, E., Murtikusuma, R. P., & Safrida, L. N. (2020). Developing mathematical reasoning problems type two-tier multiple choice for junior high school students based on ethnomathematics of jember fashion carnival. In Suratno (Ed.) *ICOLSSSTEM*. IOP Publishing. <https://doi.org/10.1088/1742-6596/1563/1/012036>.
- Andaria, M. & Hadiwinarto. (2020). Development of a two-tier multiple choice question assessment instrument to measure students science process skills on acid-base material. *ISEJ: Indonesian Science Education Journal*, 1 (3), 257-268. <https://siducat.org/index.php/isej/article/view/141>. **Coma after dot**
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi/Journal of Mathematics, Statistics, & Computing*, 9(2),95-113. <https://doi.org/10.20956/jmsk.v9i2.3402>.
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *Jurnal Suluh Pendidikan/Journal of Educational Suluh*, 17(1),32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan*[Educational evaluation basics]. Bumi Aksara.
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: the cognitive*

domain. David McKay.

Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.

Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-Tier Diagnostic Test With Certainty of Response Index on The Concepts of Fluid. In C. Anwar (Ed.), *Young Scholar Symposium on Transdisciplinary in Education and Environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf> **Incorrect alphabetical order. Move this down**

DeMars, C. E. (2010). *Item response theory*. Oxford University Press. **Move this up**

Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Hong Kong: Assessment & Support Team. <https://wlts.edb.hkedcity.net/en/home/AandLI2.html>.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87 (6), 1082–1116. <https://doi.org/10.3102/0034654317726529>.

Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12 (3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543>. **Last character is missing. Corrected link: https://doi.org/10.1111/j.1745-3992.1993.tb00543.x**

Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. **Add DOI link**

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Jamhari, M. (2021). Developing the two-tier multiple choice tests in enhancing students' higher-order thinking skills on human digestive system. *Eduproxima: Jurnal Ilmiah Pendidikan IPA*, 3 (1), 50-64. <https://doi.org/10.29100/eduproxima.v3i1.1853>.

Kartono. (2008). Equating the combined dichotomous and polytomous item test model in an

- achievement test. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>.
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how?. In A. W. Subianto (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <http://seminar.uny.ac.id/icriems/proceeding2018>.
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://doi.org/10.31327/jme.v6i2.1607>.
- Linacre, J. M. (2012). *Winstep : Rasch-model computer programs*. Winsteps.Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset Pemasaran*[Marketing Research]. Eirlangga.
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. In Cari (Ed.), *Current research in Pandemic Covid-19 Era in Indonesia - 2nd ICOSETH 2020* (pp. 44-55). University of Sebelas Maret. <https://jurnal.uns.ac.id/ijsascs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah*[Content standards for primary and secondary education]. Indonesian Government publication service.
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>.
- Prasetya, A., Rosidin, U., & Herlina, K. (2019). Development of instrument assessment for learning the polytomous response models to train higher order thinking. In C. Anwar (Ed.), *Young Scholar Symposium on Transdisciplinary in Education and Environment (YSSTEE)* (pp. 1-11). UIN Raden Intan. <http://repository.lppm.unila.ac.id/11982/1/33.%20Prasetya%202019%20J.Phys.Conf.Ser..pdf>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://doi.org/10.15294/jere.v9i2.46133>.
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha

Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>.

- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran*[Learning evaluation and assessment]. Media Akademi.
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst : Jurnal Pendidikan Fisika dan Terapannya/Karst : Journal of Physics Education and Its Application*, 4 (1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>.
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam questions for Islamic religious education and character at the elementary school level]. *An-Nahdhah/An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article/view/40>.
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Classical tes theory vs response theory? *Didaktika: Jurnal Kependidikan/Didactics: Journal of Education*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>.
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*[Rasch modeling applications in educational assessment]. Trim Komunikata.
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran*[Development of diagnostic tests in learning]. Graha Ilmu.
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9 (4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>.
- Treagust, D. (1988). Development and use of diagnostic tests to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. **Add DOI link**
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep*[Analysis of research data with rash and winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122-1?inline=1>.
- Widiyatmoko, A., & Shimizu, K. (2018). The Development of two-tier multiple choice test to assess students' conceptual understanding about light and optical instruments. *Jurnal Pendidikan IPA Indonesia*, 7 (4), 491-501. <https://doi.org/10.15294/jpii.v7i4.16591>.

Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian*[Using performance assessment for science learning and research].UPI Press.

Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>.

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :

Class/Department :

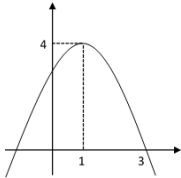
School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

<p>1. Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ C. $U_n = 4n - 1$ B. $U_n = 5n - 2$ D. $U_n = 3n + 2$ C. $U_n = 4n - 1$</p> <p>Reason:</p>	<p>2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 D. 23 B. 13 E. 24 C. 22</p> <p>Reason:</p>
<p>3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 D. 8 B. 6 E. 11 C. 7</p> <p>Reason:</p>	<p>4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 D. 344 B. 318 E. 354 C. 326</p> <p>Reason:</p>
<p>5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...</p> <p>A. 53 D. 11 B. 52 E. 10 C. 20 D. 11 E. 10</p> <p>Reason:</p>	<p>6. Given the arithmetic sequence: 4, 10, 16, 22, If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...</p> <p>A. 18 D. 24 B. 20 E. 26 C. 22</p> <p>Reason:</p>
<p>7. The nth term of an arithmetic series is $U_n = 3n - 5$. The formula for the sum of the first n terms of the series is ...</p> <p>A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$ B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$ C. $S_n = \frac{n}{2}(3n - 4)$</p> <p>Reason:</p>	<p>8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....</p> <p>A. $5n - 20$ D. $2n - 20$ B. $5n - 10$ E. $2n - 10$ C. $2n - 30$</p> <p>Reason:</p>
<p>9. The sum of all integers between 100 and 300 which are divisible by 5 is ...$S_n = \frac{n}{2}(3n - 7)$</p> <p>A. 8.200 D. 7.600 B. 8.000 E. 7.400</p>	<p>10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?</p>

<p>C. 7.800</p> <p>Reason:</p>	<p>A. 24 D. 27</p> <p>B. 25 E. 28</p> <p>C. 26</p> <p>Reason:</p>
<p>11. The middle term of an arithmetic sequence is 25. If the difference is 4 and the 5th term is 21. Then the sum of all the terms in the sequence is ...</p> <p>A. 175 D. 295</p> <p>B. 189 E. 375</p> <p>C. 275</p> <p>Reason:</p>	<p>12. A number of candies are distributed among five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces</p> <p>A. 60 D. 75</p> <p>B. 65 E. 80</p> <p>C. 70</p> <p>Reason:</p>
<p>13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...</p> <p>A. 564 D. 45</p> <p>B. 276 E. 36</p> <p>C. 48</p> <p>Reason:</p>	<p>14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...</p> <p>A. 9 D. 12</p> <p>B. 10 E. 13</p> <p>C. 11</p> <p>Reason:</p>
<p>15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...</p> <p>A. 32 D. 256</p> <p>B. 64 E. 512</p> <p>C. 128</p> <p>Reason:</p>	<p>16. The value of the middle term of the geometric sequence: 6, 3, ..., $\frac{3}{512}$ is ...</p> <p>A. $\frac{1}{16}$ D. $\frac{4}{16}$</p> <p>B. $\frac{2}{16}$ E. $\frac{5}{16}$</p> <p>C. $\frac{3}{16}$</p> <p>Reason:.....</p>
<p>17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm</p> <p>A. 18 D. 35</p> <p>B. 24 E. 40,5</p> <p>C. 27,5</p> <p>Reason:</p>	<p>18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...</p> <p>A. $\frac{3}{4}$ D. $-\frac{1}{2}$</p> <p>B. $\frac{1}{4}$ E. $-\frac{3}{4}$</p> <p>C. $\frac{1}{3}$</p> <p>Reason:</p>
<p>19. A ball falls from a height of 10 m and bounces back $\frac{3}{4}$ times its previous height. The total number of paths until the ball stops is.... m</p> <p>A. 60 D. 90</p> <p>B. 70 E. 100</p> <p>C. 80</p> <p>Reason:</p>	<p>20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.</p> <p>The value of $3a + b$ is ...</p> <p>A. 8 D. 14</p> <p>B. 10 E. 20</p> <p>C. 12</p> <p>Reason:</p>
<p>21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$.</p> <p>If $K = L$, then c is ...</p> <p>A. 12 D. 15</p>	<p>22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$.</p> <p>Then $(A + C) - (A + B)$ is ...</p> <p>A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$</p> <p>B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$</p>

<p>B. 13 C. 14 Reason:</p>	<p>E. 16 $C. \begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$ Reason:</p>
<p>23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah ... A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$ Reason:</p>	<p>24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$. If $A + B = C$, then $x + y = \dots$ A. -5 B. -1 C. 1 D. 3 E. 5 Reason:</p>
<p>25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah ... A. $\begin{bmatrix} 13 & 42 \\ 26 & 84 \end{bmatrix}$ B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$ Reason:</p>	<p>26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$ then $A(B - C) = \dots$ A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$ Reason:</p>
<p>27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ... A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$ Reason:</p>	<p>28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix. Value of x that satisfies is ... A. -5 B. -4 C. -3 D. 3 E. 4 Reason:</p>
<p>29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ... A. 0 B. 1 C. 2 D. 2 E. 4 Reason:</p>	<p>30. Transpose matrix P adalah P^T. If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ... A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$ Reason:</p>
<p>31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$. Inverse matrix AB adalah $(AB)^{-1} = \dots$ A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$</p>	<p>32. The roots of the quadratic equation $3x^2 - 4x + 3 = 0$ are ... A. $x^2 + x - 12 = 0$ B. $x^2 - x - 12 = 0$ C. $x^2 - x + 12 = 0$ D. $x^2 - 3x + 4 = 0$ E. $x^2 - 4x + 3 = 0$ Reason:</p>

Reason:	
<p>33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...</p> <p>A. -2 dan $\frac{5}{6}$ D. -2 dan $-\frac{6}{5}$</p> <p>B. 2 dan $-\frac{5}{6}$ E. -2 dan $\frac{6}{5}$</p> <p>C. 2 dan $\frac{6}{5}$</p> <p>Reason:</p>	<p>34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2. If $x_1 > x_2$, then $x_1 - x_2$ is ...</p> <p>A. -4 D. 2</p> <p>B. -2 E. 4</p> <p>C. 0</p> <p>Reason:</p>
<p>35. The roots of the quadratic equation $2x^2 - 3x - 9 = 0$ are x_1 dan x_2. Value of $x_1^2 + x_2^2$ is ...</p> <p>A. $11\frac{1}{4}$ D. $-6\frac{3}{4}$</p> <p>B. $6\frac{3}{4}$ E. $-11\frac{1}{4}$</p> <p>C. $2\frac{1}{4}$</p> <p>Reason:</p>	<p>36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α dan β. The quadratic equation $(\alpha - 2)$ dan $(\beta - 2)$ is ...</p> <p>a. $x^2 + 6x + 5 = 0$</p> <p>b. $x^2 + 6x + 7 = 0$</p> <p>c. $x^2 + 6x + 11 = 0$</p> <p>d. $x^2 - 2x + 3 = 0$</p> <p>e. $x^2 + 2x + 11 = 0$</p> <p>Reason:</p>
<p>37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...</p> <p>A. $y = x^2 - 2x + 1$</p> <p>B. $y = x^2 - 2x + 3$</p> <p>C. $y = x^2 + 2x - 1$</p> <p>D. $y = x^2 + 2x + 1$</p> <p>E. $y = x^2 + 2x + 3$</p> <p>Reason:</p>	<p>38. The figure below is a graph of the quadratic equation ? ...</p> <p>A. $y = x^2 + 2x + 3$</p> <p>B. $y = x^2 - 2x - 3$</p> <p>C. $y = -x^2 + 2x - 3$</p> <p>D. $y = -x^2 - 2x + 3$</p> <p>E. $y = -x^2 + 2x + 3$</p> <p>Reason:</p> 
<p>39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0, -4)$ then the value of $f(7)$ is ...</p> <p>A. -16 D. -19</p> <p>B. -17 E. -20</p> <p>C. -18</p> <p>Reason:</p>	<p>40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...</p> <p>A. $y = -x^2 + 2x - 3$ D. $y = -x^2 - 2x - 5$</p> <p>B. $y = -x^2 + 2x + 3$ E. $y = -x^2 - 2x + 5$</p> <p>C. $y = -x^2 - 2x + 3$</p> <p>Reason:</p>

3rd ROUND CORRECTION REPORT			
No	Reviewer Code	Reviews	Corrections made by the author
1.	R2612	The rationality of this study is not still clear and explained based on the literature	The rationale for the study has been improved and is based on the literature. Introduction (pp. 1-5)
2.	R2612	The introduction is still long.	Sentences in the introduction have been simplified (not too long). Introduction (pp. 1-5)
3.	R2612	The statements in the introduction do not serve to explain the research problem.	The statement in the introduction has been corrected to explain the research problem. The research problem 1: The test instrument developed must be accountable as a good test, and be necessary to analyze the quality of the item (Rosidin, 2017). (p. 5). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014). (p. 5). (Statement for the research problem 1: Does the open polytomous response test developed have a good category so that it can be an assessment instrument in vocational schools according to classical and modern theory?). The research problem 2: Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be known in detail... (p.3)(Statement for the research problem 2: Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools ?)

4.	R2612	Although the authors did write “Because there is still limited research on the open polytomous response test, it is necessary to conduct research.”, the strengths and weaknesses of prior studies have not been emphasized in the introduction.	<p>Reasons to conduct research: Due to the characteristics of vocational schools that place mathematics as a secondary subject and are skill-oriented, this has an impact on students' perceptions of mathematics itself, such as mathematics as an uninteresting and mechanistic subject (Putri et al., 2017), mathematics as a boring and complicated subject (Ozdemir & Onder, 2017), and mathematics as the most difficult subject (Vani et al., 2019). Because of the differences in the characteristics and perceptions of students in vocational schools with those of other schools or students, it is reasonable to conduct research by developing the open polytomous response test for students in vocational schools. (p. 4).</p>
5.	R2612	What I as to the authors, what are differences between the polytomous and two-tier tests?	<p>Differences between the polytomous and two-tier tests:</p> <p>The polytomous test: Multiple-choice tests with only two answer choices are called "dichotomous tests," and multiple-choice tests with more than two answer choices are called "polytomous tests" (Kartono, 2008). (p. 2)</p> <p>The two-tier tests: In the 80s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). (p. 3).</p>
6..	R2612	In the discussion, remove these words “..do not be in a hurry to revise..”	<p>Discussion: The sentence "... do not be in a hurry to revise..." has been removed, and replaced with the sentence: “That is, an item that is not in a good category with classical theory should be analyzed according to modern theory before revising or replacing the item” (p. 25).</p>
7.	R2612	The discussion has too few references to discuss the results. Also, it is too limited.	<p>Discussion has been added (reference) The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas, and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use</p>

			<p>both patterns can answer the questions correctly as shown in Figure 15 below.</p> <div data-bbox="1108 263 1630 534"> <p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p> </div> <div data-bbox="1653 263 1989 534"> <p style="text-align: center;">(i)</p> </div> <div data-bbox="2000 263 2235 534"> </div> <p><i>Figure 15. Student Answer Patterns with (i) Formulas, and (ii) Trial and Error</i></p> <p>Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, use a drawing or model, analogy, and formula. Syahlan (2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syahlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.</p> <p>The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching a material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know the initial abilities of students are that they can develop professionally centered on students (Gonzalez, 2018), adjust the level of cognitive engagement and increase student learning engagement Dong et al., 2020), assist teachers in designing pedagogical practices, and correct students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if she knows the students' actual</p>
--	--	--	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

			abilities, and this can be known by the teacher if she uses the open polytomus response test. (pp. 26-27)
8.	R2612	The problems with the use of the English language are continuing. This paper needs copyediting by a native speaker and an expert in the field.	English has been consulted by experts in the field

The Development of an Assessment Instrument Using Polytomous Response in Mathematics

Abstract: This research is development research that aims to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. This research design uses the Plomp model, which consists of five stages, namely: preliminary investigation, design, realization or construction, trial, revision, and implementation (testing). The study was conducted in three vocational schools in Lampung, Indonesia. The study involves 413 students, consisting of 191 male students and 222 female students. The data was collected through questionnaires and tests. The questionnaire was to identify instruments commonly used by teachers and to validate them by experts. The test used multiple-choice tests with open reasons for as many as 40 items. The data was analyzed using both classical and modern theories. The results show that the open polytomous response test has a good category according to classical and modern theory, and the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: assessment instrument, classical and modern theory, vocational school, the polytomous responses

Introduction

Assessment is an important activity and needs to be done by teachers in schools. The conventional paradigm often interprets the assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syarifuddin, 2020). Referring to the current paradigm, assessment in schools is divided into three types: assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). These three types of assessments aim to

provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013), as shown in the following assessment pyramid (Figure 1).



Figure 1. Assessement Pyramid

Assessment can be used with tests. A test is a tool or procedure used to find out or measure students' abilities in certain areas with certain rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides an answer choice, and one of the choices is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has strengths or weaknesses compared to each other. The strength of multiple-choice tests over essays is that multiple-choice tests can be conducted for many students, are more objective, and the test results can be known more quickly, but they have a weakness, namely that the multiple-choice test is not able to see the actual abilities of students and the answers tend to be guesses or tried out (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (getting a score of 1 for the correct answer, and a score of 0 for the wrong answer choice). Multiple-choice tests with only two answer choices are called "dichotomous tests," and multiple-choice tests with more than two answer choices are called "polytomous tests" (Kartono, 2008).

Until now, multiple-choice tests have been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice tests, in the last four decades, experts have developed multiple-choice tests by combining multiple-choice tests and essays into multiple-choice tests with

reasons and are called the polytomous response test (Suwanto, 2012). The polytomous response test score is 1 - 4. Score of 4 for the correct answer and reason, score of 3 for the correct answer but the wrong reason, score of 2 for the wrong answer but the correct reason, and score of 1 for the wrong answer and reason (Kartono, 2008).

In the 80s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as tests on calculus material (Khiyarunnisa & Retnawati, 2018), tests on acid and base materials (Andaria & Hadiwinarto, 2020), tests on human digestive system material (Jamhari, 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be known in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easy to observe (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test to be the open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are tests on calculus material in universities (Yang et al., 2017) and tests on physics

material in high schools (Prasetya et al., 2019). Both of these studies have developed the open polytomous response test for students in college or high school. Students in colleges and high schools place mathematics as a primary subject, while students in vocational schools place mathematics as a secondary subject (Oktaria, 2016). In addition, graduate students in vocational schools are more oriented towards practical abilities and skills, in contrast to graduate students in colleges or high schools, who are academically oriented (Permendikbud, 2016). Due to the characteristics of vocational schools that place mathematics as a secondary subject and are skill-oriented, this has an impact on students' perceptions of mathematics itself, such as mathematics as an uninteresting and mechanistic subject (Putri et al., 2017), mathematics as a boring and complicated subject (Ozdemir & Onder, 2017), and mathematics as the most difficult subject (Vani et al., 2019). Because of the differences in the characteristics and perceptions of students in vocational schools with those of other schools or students, it is reasonable to conduct research by developing the open polytomous response test for students in vocational schools.

The test instrument developed must be accountable as a good test, and be necessary to analyze the quality of the item (Rosidin, 2017). There are two theories for analyzing item quality, namely classical and modern. Classical theory is a measurement theory for assessing tests based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities and is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because

in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

This research is development research that aims to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problem is stated as follows: (1) does the open polytomous response test developed have a good category so that it can be an assessment instrument in vocational schools according to classical and modern theory?, and (2) does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools ?

Methodology

Research Design

This research is a research and development that refers to Plomp's model (Plomp, 2013), with the research procedure consisting of five stages, namely preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

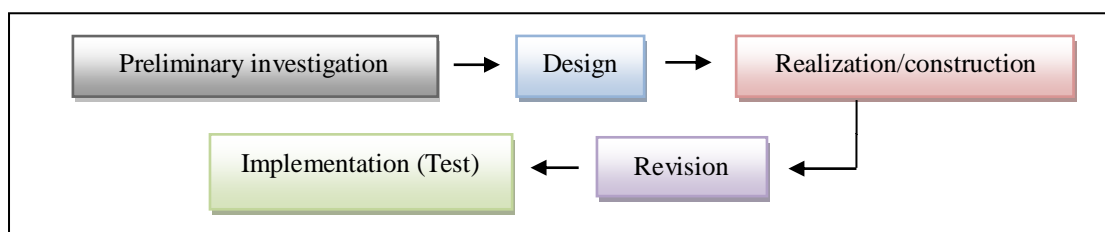


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the assessment instruments used by teachers so far. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make an expert assessment questionnaire sheet. The realization/construction stage is developing the items test and also the expert validation process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and analyze the results of the test.

Research Subject

The subjects of the study are students at a vocational school in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools are three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects are 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data was collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two people who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert are the suitability of the items with the indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining content validity, the instrument was tested on students. Then, it was continued by determining the validity of the construct and its reliability, with the aim of ensuring that the instrument can be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contains five answer choices along with the reasons. Student scores refer to the polytomous score in the Partial Credit Model, where answer choices and reasons are related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected research data is analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. *Questionnaire data analysis (qualitative analysis)*

There are two sets of questionnaire data, namely: identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically for expert judgment, it is continued with an analysis of expert agreement that uses the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. *Test data analysis (empirical analysis)*

After the content validity test, it is followed by the construct validity and reliability tests. The construct validation test uses exploratory factor analysis. The instrument is considered to have a good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5

(Retnawati, 2014). Reliability test using Cronbach's alpha formula The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further tests can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the IteMan program is used for classical theory and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program is used because it has several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3–0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is

close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.

- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigen values of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test uses the Eigen value analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information

function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it is found that so far, the teacher has never used the polytomous response. As many as 80% of teachers use essay tests and 20% of teachers use multiple-choice tests, with each instrument consisting of 2–5 items. In addition, about 10% of teachers use this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who do not use assessment as an improvement in learning are caused by several aspects, such as: teachers do not understand assessment (20%), teachers do not know how to analyze assessments (50%), and teachers do not know how to develop good assessment questions (30%). The following is a summary of the questionnaire from the identification data.

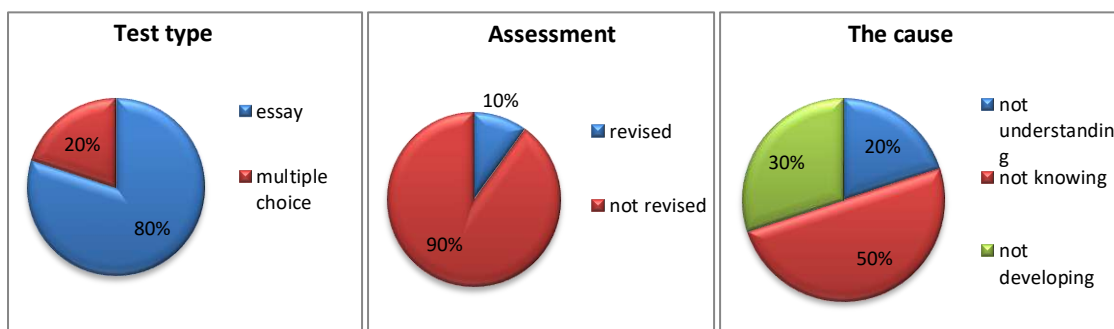


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments show that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement is obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provide some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it is followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's Alpha coefficient value of 0.89 (more than 0.6). It means that the instrument has good reliability so

that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way does not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items can be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised
4	0.540	good	0.304	Good	24	0.438	good	-0.071	revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	revised

Based on Table 7, it is found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are included in the good category. Meanwhile, only two items on discrimination have good categories, and the remaining items need to be revised. The results indicate that all items are good based on the level of difficulty, but almost all items need to be revised for item discrimination.

The Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the eigenvalues. The eigenvalue is then used to calculate the percentage of explained variance, as well as describe the scree plot (Retnawati, 2014). The output of factor analysis is the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigen values and scree plot analysis results. If the cumulative percentage of the first factor eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first factor Eigen values is 20.220%. Because the Eigen value is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigen Values		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigen values.

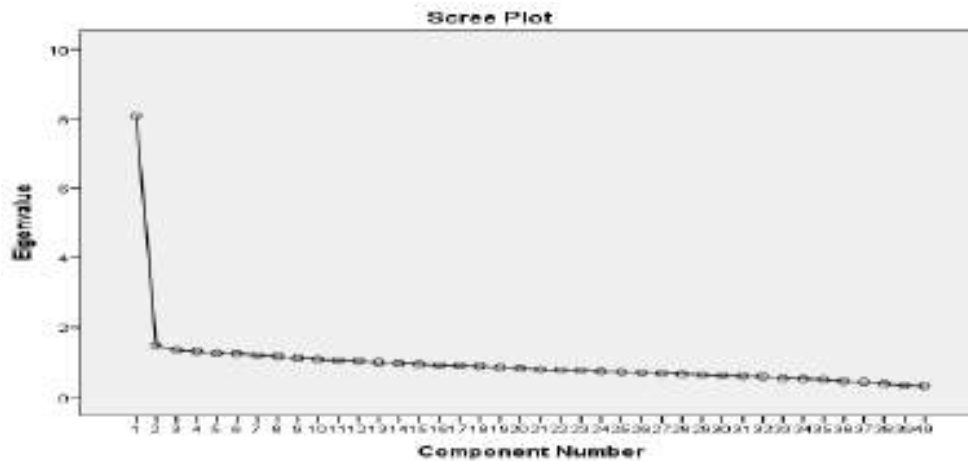


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the eigenvalues immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test is analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Corr is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items match the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	TOTAL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.43	.07	.97	-.4	.98	-.5	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.3	.97	-.5	.43	.44	48.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	988	413	-.03	.07	.97	-.4	.98	-.4	.39	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.3	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.31	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.3	1.08	1.3	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.0	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	48.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.28	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	828	413	.72	.07	.83	-8.3	.88	-6.2	.39	.43	38.1	48.6	Q39
40	801	413	.84	.07	.74	-4.3	.74	-4.6	.34	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level is analyzed using the Winsteps program, and the results obtained can be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-0.5	.97	-0.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-0.4	.98	-0.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-0.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-0.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-0.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-0.7	.96	-0.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-0.08	.07	.97	-0.4	.97	-0.4	.47	.44	50.8	50.9	Q20
21	1010	413	-0.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-0.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-0.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-0.4	.97	-0.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-0.01	.07	.97	-0.5	.97	-0.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-0.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-0.6	.96	-0.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-0.1	1.00	-0.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

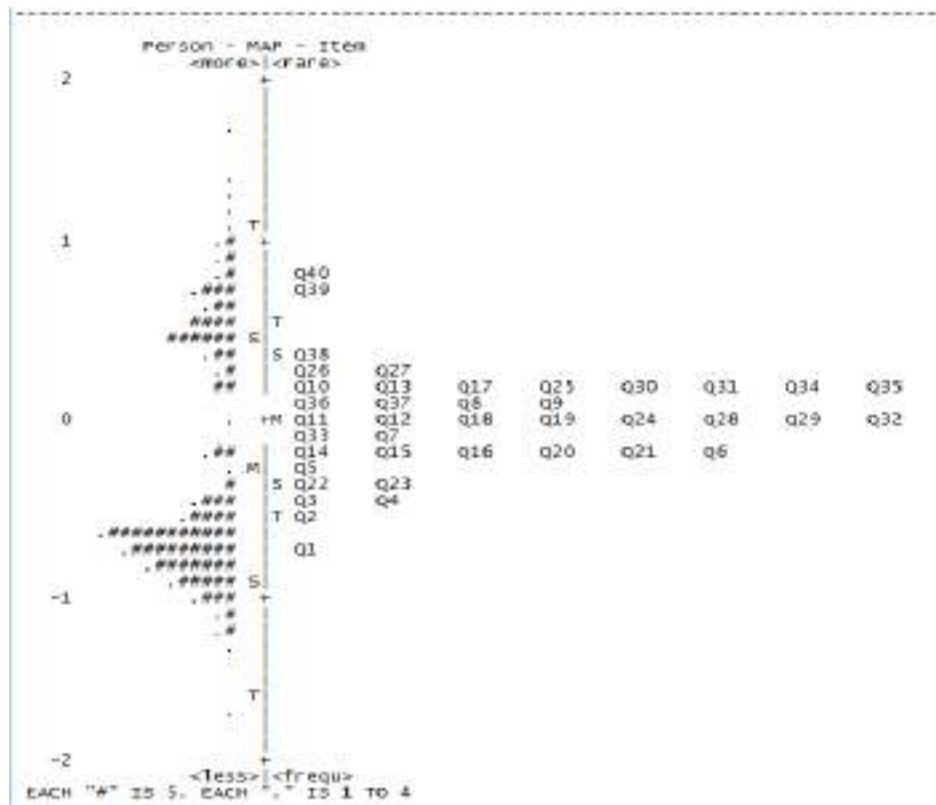


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (column PT-Measure Corr). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than using classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen

that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability is measured by using a test that is expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

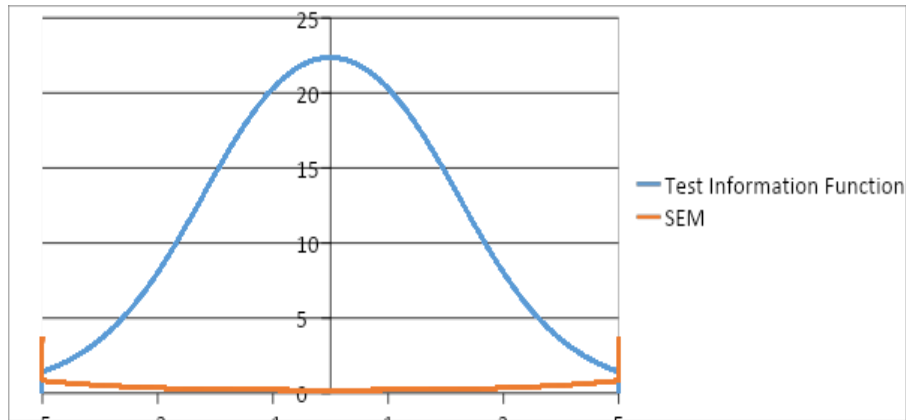


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function

expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in SMK is based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

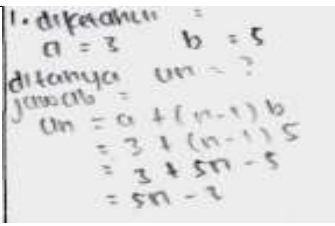
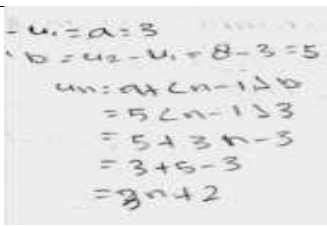
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	 <p>1. diketahui $a = 3$ $b = 5$ ditanya $u_n = ?$ jawab $u_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	 <p>$u_1 = a = 3$ $b = u_2 - u_1 = 8 - 3 = 5$ $u_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>

Figure 10. Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only

writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

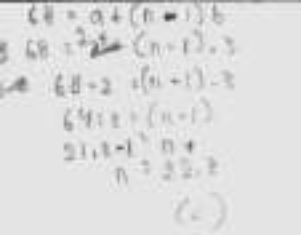
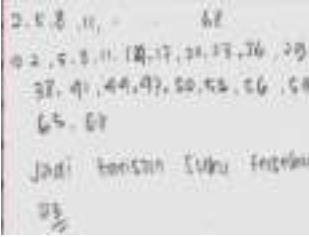
Question 2:	Pattern 1	Pattern 2
<p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.</p> <p>The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>		

Figure 11. Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

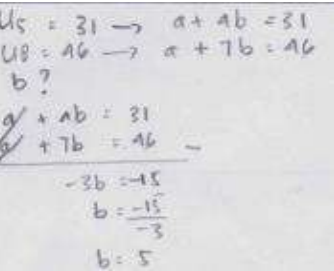
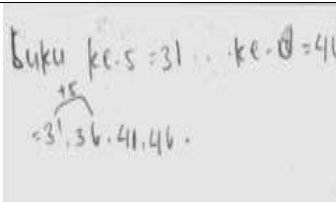
Question 3:	Pattern 1	Pattern 2
<p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>		

Figure 12. Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the n th term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the n th term.

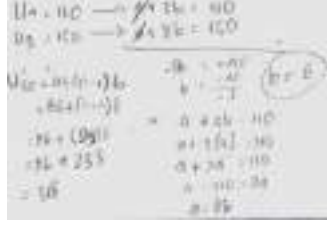
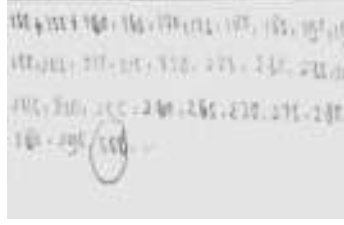
Question 4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>	 <p>Handwritten work for Pattern 1: $U_4 = 110 \rightarrow a + 3d = 110$, $U_9 = 150 \rightarrow a + 8d = 150$. Solving for d gives $d = 10$, then $a = 70$. The 30th term is $U_{30} = 70 + 29(10) = 360$.</p>	 <p>Handwritten work for Pattern 2: Lists terms starting from 110 and 150, continuing the sequence until the 30th term, which is 360.</p>

Figure 13. Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

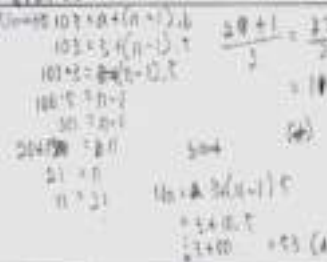
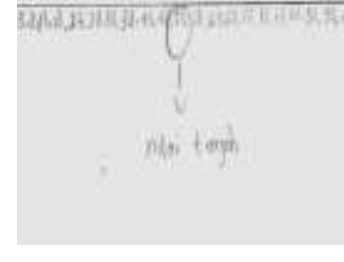
Question 5:	Answer Pattern 1	Answer Pattern 2
<p>An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...308</p> <p>A. 53 B. 52 C. 20 D. 11 E. 10</p> <p>Reason:</p>	 <p>Handwritten work for Answer Pattern 1: $U_n = 3 + (n-1)5 = 103$, solving for n gives $n = 21$. The middle term is $U_{11} = 3 + 10(5) = 53$.</p>	 <p>Handwritten work for Answer Pattern 2: Lists terms of the sequence: 3, 8, 13, 18, ..., 103. The middle term is identified as 53.</p>

Figure 14. Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice tests are easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score, then the result is multiplied by 10 to get a value range of 0–10, or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8, 56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the

teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomus response test is more secondhand and complete than other assessment instruments, such as multiple choice tests (Gierl et al., 2017) or essay tests (Putri et al., 2020).

Discussion

This research is development research that aims to develop a good mathematical assessment instrument using polytomic responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). **That is, an item that is not in a good category with classical theory should be analyzed according to modern theory before revising or replacing the item.**

The study of assessment with an open response polytomus test was carried out by Yang et al. (2017). However, this Yang study has several fundamental differences, namely the research objectives and data analysis. The aim of this research is to diagnose the errors of university students in the concept of calculus, not to produce a good assessment instrument. The data analysis used parametric statistics (covariance), not item analysis (classic and modern). Because the objectives and data analysis are different, the results of this study cannot be compared with the results of any other study. However, this research has provided other

researchers with information in developing tests, such as the suitability of items with indicators, use of language, and preparation of answer choices.

There are other studies related to the response politomus test, namely Sarea (2018) and Saepuzaman (2021). Sarea's research states that the response polytomus test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomus test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas, and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 15 below.

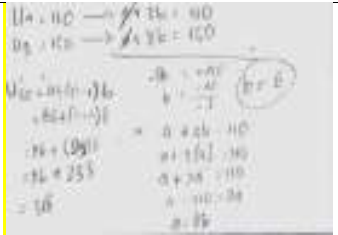
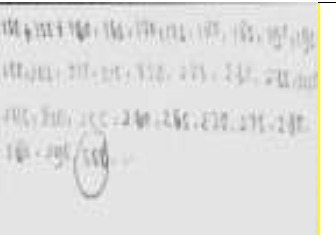
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>	 <p style="text-align: center;">(i)</p>	 <p style="text-align: center;">(ii)</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------

Figure 15. Student Answer Patterns with (i) Formulas, and (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, use a drawing or model, analogy, and formula. Syahlan (2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas

methods to solve difficult questions. If combined with student answers and Syahlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching a material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, conclusions are obtained, namely that (1) the open polytomous response test has a good category according to classical and modern theory. So, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices.

Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Andaria, M., & Hadiwinarto. (2020). Development of a two-tier multiple choice question assessment instrument to measure students' science process skills on acid-base material. *ISEJ: Indonesian Science Education Journal*, 1 (3), 257-268. <https://siducat.org/index.php/isej/article/view/141>.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi/ Journal of Mathematics, Statistics, & Computing*, 9(2), 95-113. <https://doi.org/10.20956/jmsk.v9i2.3402>.
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *Jurnal Suluh Pendidikan/Journal of Educational Suluh*, 17(1), 32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: the cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. E. (2010). *Item response theory*. Oxford University Press.
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-Tier Diagnostic Test With Certainty of Response Index on The Concepts of Fluid. In C. Anwar (Ed.), *Young Scholar Symposium on Transdisciplinary in Education and Environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf>
- Dong, A., Jong, M. S., & King, R. B. (2020). How Does Prior Knowledge Influence Learning Engagement? The Mediating Roles of Cognitive Load and Help-Seeking. *Frontier Psychology Journal*. <https://doi.org/10.3389/fpsyg.2020.591203>

- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Hong Kong: Assessment & Support Team. <https://wlts.edb.hkedcity.net/en/home/AandLI2.html>.
- Geofrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. *Researchgate Journal*. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87 (6), 1082–1116. <https://doi.org/10.3102/0034654317726529>.
- Gonzalez, G. (2018). Understanding Teacher Noticing of Students' Prior Knowledge: Challenges and Possibilities. *The Mathematics Enthusiast*, 15 (3), 483-528. <https://doi.org/10.54870/1551-3440.1442>
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12 (3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Jamhari, M. (2021). Developing the two-tier multiple choice tests in enhancing students' higher-order thinking skills on human digestive system. *Eduproxima: Jurnal Ilmiah Pendidikan IPA*, 3 (1), 50-64. <https://doi.org/10.29100/eduproxima.v3i1.1853>.
- Kartono. (2008). Equating the combined dichotomous and polytomuos item test model in an achievement test. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>.
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how?. In A. W. Subianto (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <http://seminar.uny.ac.id/icriems/proceeding2018>.
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation

- tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://doi.org/10.31327/jme.v6i2.1607>.
- Linacre, J. M. (2012). *Winstep : Rasch-model computer programs*. Winsteps. Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Eirlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically (Second edition)*. Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction* , 10 (2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. In Cari (Ed.), *Current research in Pandemic Covid-19 Era in Indonesia - 2nd ICOSETH 2020* (pp. 44-55). University of Sebelas Maret. <https://jurnal.uns.ac.id/ijsascs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Ozdemir, H., & Onder, N. (2017). Vocational high school students' perceptions of success in mathematics. *International Electronic Journal of Mathematics Education* , 12 (3), 493-502. <https://doi.org/10.29333/iejme/627>.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Content standards for primary and secondary education]. Indonesian Government publication service.
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>.
- Prasetya, A., Rosidin, U., & Herlina, K. (2019). Development of instrument sssessment for learning the polytomous response models to train higher order thinking. In C. Anwar (Ed.), *Young Scholar Symposium on Trandisciplinaty in Education and Environment (YSSTEE)* (pp. 1-11). UIN Raden Intan. <http://repository.lppm.unila.ac.id/11982/1/33.%20Prasetya%202019.J.Phys.Conf.Ser..pdf>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://doi.org/10.15294/jere.v9i2.46133>.
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6 1), 97-107. <https://doi.org/10.15294/ujme.v6i1.12643>.

- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst : Jurnal Pendidikan Fisika dan Terapannya/ Karst : Journal of Physics Education and Its Application*, 4 (1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>.
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam questions for Islamic religious education and character at the elementary school level]. *An-Nahdhah/An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article/view/40>.
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Classical tes theory vs response theory? *Didaktika: Jurnal Kependidikan/Didactics: Journal of Education*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>.
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunika.
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic tests in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4 (6), 358-369. <http://idealmathedu.p4tkmatematika.org>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9 (4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>.
- Treagust, D. (1988). Development and use of diagnostic tests to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep* [Analysis of research data with rash and winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122-1?inline=1>.

- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako/ Electronic Journal of Tadulako Mathematics Education*, 6(4), 455-468. <http://jurnal.untad.ac.id/jurnal/index.php/JEPMT/article/view/14181>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8 (2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>.

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :
Class/Department :
School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

- Given an arithmetic sequence: 3, 8, 13, 18,
The formula for the nth term of the sequence is
A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
C. $U_n = 4n - 1$
Reason:
- Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.
The number of terms in the sequence is...
A. 12 D. 23
B. 13 E. 24
C. 22
Reason:
- An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
A. 5 D. 8
B. 6 E. 11
C. 7
Reason:
- The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
A. 308 D. 344
B. 318 E. 354
C. 326
Reason:
- An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
Then the middle term of the sequence is ...
A. 53 D. 11
B. 52 E. 10
C. 20
D. 11
E. 10
Reason:
- Given the arithmetic sequence: 4, 10, 16, 22,
If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
A. 18 D. 24
B. 20 E. 26
C. 22
Reason:
- The nth term of an arithmetic series is $U_n = 3n - 5$.
The formula for the sum of the first n terms of the series is ...
A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
C. $S_n = \frac{n}{2}(3n - 4)$
Reason:
- The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....
A. $5n - 20$ D. $2n - 20$
B. $5n - 10$ E. $2n - 10$
C. $2n - 30$
Reason:
- The sum of all integers between 100 and 300 which are divisible by 5 is ... $S_n = \frac{n}{2}(3n - 7)$
A. 8.200 D. 7.600
B. 8.000 E. 7.400
C. 7.800
Reason:
- PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
A. 24 D. 27
B. 25 E. 28
C. 26
Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah
- ...
- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
- B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
- C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$
- Reason:
24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.
- If $A + B = C$, then $x + y = \dots$
- A. -5 D. 3
B. -1 E. 5
C. 1
- Reason:
25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah
- ...
- A. $\begin{bmatrix} 13 & 42 \\ 26 & 84 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \\ 30 & 36 \end{bmatrix}$
B. $\begin{bmatrix} 26 & 84 \\ 26 & 42 \end{bmatrix}$ E. $\begin{bmatrix} 13 & 84 \\ 30 & 36 \end{bmatrix}$
C. $\begin{bmatrix} 26 & 42 \\ 26 & 42 \end{bmatrix}$
- Reason:
26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$ then $A(B - C) = \dots$
- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$
- Reason:
27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...
- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$
- Reason:
28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix. Value of x that satisfies is ...
- A. -5 D. 3
B. -4 E. 4
C. -3
- Reason:
29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ...
- A. 0 D. 2
B. 1 E. 4
C. 2
- Reason:
30. Transpose matrix P adalah P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ...
- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$
- Reason:
31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$. Inverse matrix AB adalah $(AB)^{-1} = \dots$
- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$
- Reason:
32. The roots of the quadratic equation $3x^2 - 4x - 4 = 0$ are ...
- A. $x^2 + x - 12 = 0$
B. $x^2 - x - 12 = 0$
C. $x^2 - x + 12 = 0$
D. $x^2 - 3x + 4 = 0$
E. $x^2 - 4x + 3 = 0$
- Reason:
33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...
- A. -2 dan $\frac{5}{6}$ D. -2 dan $-\frac{6}{5}$
B. 2 dan $-\frac{5}{6}$ E. -2 dan $\frac{6}{5}$
C. 2 dan $\frac{6}{5}$
- Reason:
34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...
- A. -4 D. 2
B. -2 E. 4
C. 0
- Reason:

35. The roots of the quadratic equation $2x^2 - 3x - 9 = 0$ are x_1 dan x_2 . Value of $x_1^2 + x_2^2$ is ...

- A. $11\frac{1}{4}$
- B. $6\frac{3}{4}$
- C. $2\frac{1}{4}$
- D. $-6\frac{3}{4}$
- E. $-11\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...

- A. $y = x^2 - 2x + 1$
- B. $y = x^2 - 2x + 3$
- C. $y = x^2 + 2x - 1$
- D. $y = x^2 + 2x + 1$
- E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0, -4)$ then the value of $f(7)$ is ...

- A. -16
- B. -17
- C. -18
- D. -19
- E. -20

Reason:

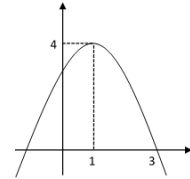
36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α dan β . The quadratic equation $(\alpha - 2)$ dan $(\beta - 2)$ is ...

- a. $x^2 + 6x + 5 = 0$
- b. $x^2 + 6x + 7 = 0$
- c. $x^2 + 6x + 11 = 0$
- d. $x^2 - 2x + 3 = 0$
- e. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation ? ...

- A. $y = x^2 + 2x + 3$
- B. $y = x^2 - 2x - 3$
- C. $y = -x^2 + 2x - 3$
- D. $y = -x^2 - 2x + 3$
- E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...

- A. $y = -x^2 + 2x - 3$
- B. $y = -x^2 + 2x + 3$
- C. $y = -x^2 - 2x + 3$
- D. $y = -x^2 - 2x - 5$
- E. $y = -x^2 - 2x + 5$

Reason:

To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Dear Dr. Sutiarso,

We have received your 3rd revised paper We have sent it to our reviewers again in order to check. We will inform you when we get the result from our reviewers.

If the reviewers confirm your revised paper, we will send the acceptance letter to you.

Thank you for your patience.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

[Quoted text hidden]



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

4th round corrections request for the manuscript ID# 21112502244011

3 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>
Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Mon, Apr 4, 2022 at 3:48 PM

Dear Dr. Sutiarso,

Please see the attached file as the second round corrections.

Please remove the old highlights and re-highlight for new edited parts. We don't need a new correction report.

We are looking forward to getting your second revised paper until **April 11, 2022**.**PS. If the all corrections can't be done, the editorial process will be cancelled.**

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com

On 29-Mar-22 12:25 PM, SUGENG SUTJARSO wrote:

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.

I have revised the article according to the reviewer's suggestion. Here I attach (1) 3rd round of article revisions,
and (2) 3rd correction report.

Best regards,

Sugeng Sutiarso
Lampung University

On Wed, Mar 16, 2022 at 7:00 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarso,

Please see the attached file as the second round corrections.

Please remove the old highlights and re-highlight for new edited parts. We need a new correction report.

We are looking forward to getting your second revised paper until **March 30, 2022**.**PS. If the all corrections can't be done, the editorial process will be cancelled.**

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com

On 15-Mar-22 3:14 PM, SUGENG SUTJARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research.

I apologize for my error in citation (there is a problem in my computer). Here I re-send my article.

Thank you for this opportunity to improve.

Best regards,
Sugeng Sutiarmo
Lampung University, Indonesia.

On Tue, Mar 15, 2022 at 1:57 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

In-text citations are not visible in the edited file (we guess it's because of the program you were using). Could you correct the citations and re-send it please urgently?

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 13-Mar-22 3:51 PM, SUGENG SUTIARSO wrote:

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.

I have revised the second round of corrections according to the reviewer's suggestion. Here I attach a correction report and revised article.

Thank you.

Best regards,

Sugeng Sutiarmo
Lampung University

On Mon, Feb 28, 2022 at 7:59 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarmo,

Please see the attached file as the second round corrections.

Please remove the old highlights and re-highlight for new edited parts. We need a new correction report.

We are looking forward to getting your second revised paper until **March 14, 2022**.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com

On 22-Feb-22 4:48 PM, SUGENG SUTIARSO wrote:


Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.


I have revised the article according to the reviewer's suggestion. Here I attach (1) a revised article, (2) a correction report, and (3) a proofreading certificate from my university's language center.

Best regards,

Sugeng Sutiarmo
Lampung University

2 attachments

 **4TH ROUND_EU-JER_21112502244011_R2613.docx**
1351K

 **4TH ROUND_EU-JER_21112502244011_R2612.docx**
137K

SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Mon, Apr 11, 2022 at 9:35 PM

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.

I have revised the article according to the reviewer's suggestion. Here I attach 4th round of article revisions.

Best regards,

Sugeng Sutiarso
Lampung University

[Quoted text hidden]

 **4th Revision_Article Sugeng Sutiarso et al.docx**
1704K

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>

Tue, Apr 12, 2022 at 3:03 AM

Dear Dr. Sutiarso,

We have received your 4th revised paper We have sent it to our reviewers again in order to check. We will inform you when we get the result from our reviewers.

If the reviewers confirm your revised paper, we will send the acceptance letter to you.

Thank you for your patience.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

[Quoted text hidden]



European Journal of Educational Research

ISSN: 2165-8714

<http://www.eu-jer.com/>

Review Form

Manuscript ID:	EU- JER_ID#_21112502244011	Date: 3 April 2022
Manuscript Title:	The Development of an Assessment Instrument Using Polytomous Response in Mathematics	

ABOUT MANUSCRIPT (Mark with "X" one of the options)	Accept	Weak	Refuse	Not Available
Language is clear and correct		X		
Literature is well written		X		
References are cited as directed by APA	X			
The research topic is significant to the field	X			
The article is complete, well organized and clearly written		X		
Research design and method is appropriate	X			
Analyses are appropriate to the research question	X			
Results are clearly presented	X			
A reasonable discussion of the results is presented		X		
Conclusions are clearly stated	X			
Recommendations are clearly stated	X			

GENERAL REMARKS AND RECOMMENDATIONS TO THE AUTHOR

The statements in the introduction are still explaining the research problem. For example,

“Several studies on the closed polytomous response test have been carried out, such as tests on calculus material (Khiyarunnisa & Retnawati, 2018), tests on acid and base materials (Andaria & Hadiwinarto, 2020), tests on human digestive system material (Jamhari, 2021), and test on mathematical connection material”

These sentences are not directly related to mathematics. While explaining the research problem, please keep focusing only on studies conducted on mathematics.

Also, I think the followings sentences are too superficial to explain the strengths and weaknesses of prior studies.

“Due to the characteristics of vocational schools that place mathematics as a secondary subject and are skill-oriented, this has an impact on students' perceptions of mathematics itself, such as mathematics as an uninteresting and mechanistic subject (Putri et al., 2017), mathematics as a boring and complicated subject (Ozdemir & Onder, 2017), and mathematics as the most difficult subject (Vani et al., 2019). Because of the differences in the characteristics and perceptions of students in vocational schools from those of other schools or students, it is reasonable to conduct research by developing the open polytomous response test for students in vocational schools. (p. 4).”

It is not clear what are the studies of prior studies that developed a polytomous response test in the vocational education context. What are the strengths and weaknesses of these studies?

The discussion is not still a real discussion. The differences and similarities between previous studies and



European Journal of Educational Research

ISSN: 2165-8714

<http://www.eu-jer.com/>

this study are not still understandable.

Language

The language and organization of the paper are not well constructed. The manuscript strongly needs proofreading by a native speaker. The overall writing is poor. The text still includes most grammatical and logical errors.

Sincerely,

THE DECISION (Mark with "X" one of the options)

Accepted: Correction not required	
Accepted: Minor correction required	
Conditionally Accepted: Major Correction Required (Need second review after corrections)	X
Refused	

Reviewer Code: R2612 (The name of referee is hidden because of blind review)

The Development of an Assessment Instrument Using Polytomous Response in Mathematics

Abstract: This research is development research that aims to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. This research design uses the Plomp model, which consists of five stages, namely: preliminary investigation, design, realization or construction, trial, revision, and implementation (testing). The study was conducted in three vocational schools in Lampung, Indonesia. The study involves 413 students, consisting of 191 male students and 222 female students. The data was collected through questionnaires and tests. The questionnaire was to identify instruments commonly used by teachers and to validate them by experts. The test used multiple-choice tests with open reasons for as many as 40 items. The data was analyzed using both classical and modern theories. The results show that the open polytomous response test has a good category according to classical and modern theory, and the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: assessment instrument, classical and modern theory, vocational school, the polytomous responses

Introduction

Assessment is an important activity and needs to be done by teachers in schools. The conventional paradigm often interprets the assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syarifuddin, 2020). Referring to the current paradigm, assessment in schools is divided into three types: assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). These three types of assessments aim to

provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013), as shown in the following assessment pyramid (Figure 1).



Figure 1. Assesment Pyramid

Assessment can be used with tests. A test is a tool or procedure used to find out or measure students' abilities in certain areas with certain rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides an answer choice, and one of the choices is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has strengths or weaknesses compared to each other. The strength of multiple-choice tests over essays is that multiple-choice tests can be conducted for many students, are more objective, and the test results can be known more quickly, but they have a weakness, namely that the multiple-choice test is not able to see the actual abilities of students and the answers tend to be guesses or tried out (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (getting a score of 1 for the correct answer, and a score of 0 for the wrong answer choice). Multiple-choice tests with only two answer choices are called "dichotomous tests," and multiple-choice tests with more than two answer choices are called "polytomous tests" (Kartono, 2008).

Until now, multiple-choice tests have been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice tests, in the last four decades, experts have developed multiple-choice tests by combining multiple-choice tests and essays into multiple-choice tests with

reasons and are called the polytomous response test(Suwarto, 2012). The polytomous response test score is 1 - 4. Score of 4 for the correct answer and reason, score of 3 for the correct answer but the wrong reason, score of 2 for the wrong answer but the correct reason, and score of 1 for the wrong answer and reason(Kartono, 2008).

In the 80s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test(Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as tests on calculus material (Khiyarunnisa & Retnawati, 2018), tests on acid and base materials (Andaria & Hadiwinarto, 2020), tests on human digestive system material (Jamhari, 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be known in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easy to observe (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test to be the open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are tests on calculus material in universities (Yang et al., 2017) and tests on physics

material in high schools (Prasetya et al., 2019). Both of these studies have developed the open polytomous response test for students in college or high school. Students in colleges and high schools place mathematics as a primary subject, while students in vocational schools place mathematics as a secondary subject (Oktaria, 2016). In addition, graduate students in vocational schools are more oriented towards practical abilities and skills, in contrast to graduate students in colleges or high schools, who are academically oriented (Permendikbud, 2016). Due to the characteristics of vocational schools that place mathematics as a secondary subject and are skill-oriented, this has an impact on students' perceptions of mathematics itself, such as mathematics as an uninteresting and mechanistic subject (Putri et al., 2017), mathematics as a boring and complicated subject (Ozdemir & Onder, 2017), and mathematics as the most difficult subject (Vani et al., 2019). Because of the differences in the characteristics and perceptions of students in vocational schools with those of other schools or students, it is reasonable to conduct research by developing the open polytomous response test for students in vocational schools.

The test instrument developed must be accountable as a good test, and be necessary to analyze the quality of the item (Rosidin, 2017). There are two theories for analyzing item quality, namely classical and modern. Classical theory is a measurement theory for assessing tests based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities and is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because

in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

This research is development research that aims to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problem is stated as follows:(1) does the open polytomous response test developed have a good category so that it can be an assessment instrument in vocational schools according to classical and modern theory?, and (2) does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools ?

Methodology

Research Design

This research is a research and development that refers to Plomp's model (Plomp, 2013), with the research procedure consisting of five stages, namely preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

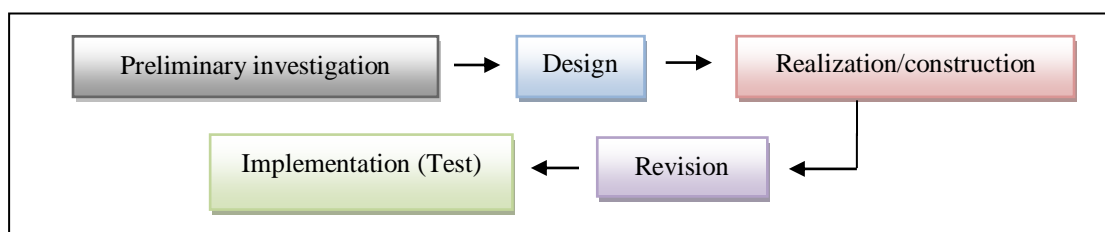


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the assessment instruments used by teachers so far. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make an expert assessment questionnaire sheet. The realization/construction stage is developing the items test and also the expert validation process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and analyze the results of the test.

Research Subject

The subjects of the study are students at a vocational school in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools are three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects are 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data was collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two people who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert are the suitability of the items with the indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining content validity, the instrument was tested on students. Then, it was continued by determining the validity of the construct and its reliability, with the aim of ensuring that the instrument can be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contains five answer choices along with the reasons. Student scores refer to the polytomous score in the Partial Credit Model, where answer choices and reasons are related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected research data is analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. *Questionnaire data analysis (qualitative analysis)*

There are two sets of questionnaire data, namely: identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically for expert judgment, it is continued with an analysis of expert agreement that uses the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. *Test data analysis (empirical analysis)*

After the content validity test, it is followed by the construct validity and reliability tests. The construct validation test uses exploratory factor analysis. The instrument is considered to have a good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5

(Retnawati, 2014). Reliability test using Cronbach's alpha formula The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60(Arikunto, 2012). If the instrument has good construct validation, further tests can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the IteMan program is used for classical theory and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program is used because it has several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3–0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is

close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.

- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigen values of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test uses the Eigen value analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information

function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it is found that so far, the teacher has never used the polytomous response. As many as 80% of teachers use essay tests and 20% of teachers use multiple-choice tests, with each instrument consisting of 2–5 items. In addition, about 10% of teachers use this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who do not use assessment as an improvement in learning are caused by several aspects, such as: teachers do not understand assessment (20%), teachers do not know how to analyze assessments (50%), and teachers do not know how to develop good assessment questions (30%). The following is a summary of the questionnaire from the identification data.

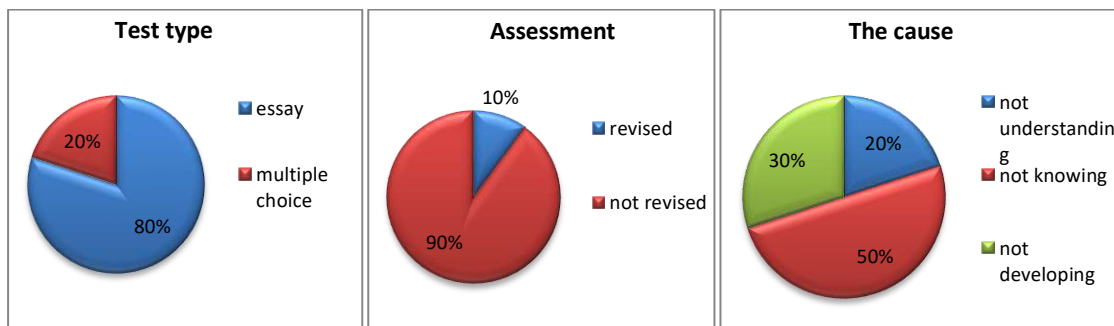


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments show that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement is obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provide some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it is followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's Alpha coefficient value of 0.89 (more than 0.6). It means that the instrument has good reliability so

that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way does not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items can be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised
4	0.540	good	0.304	Good	24	0.438	good	-0.071	revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	revised

Based on Table 7, it is found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are included in the good category. Meanwhile, only two items on discrimination have good categories, and the remaining items need to be revised. The results indicate that all items are good based on the level of difficulty, but almost all items need to be revised for item discrimination.

Analysis of Test Data with Modern Theory

The Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the eigenvalues. The eigenvalue is then used to calculate the percentage of explained variance, as well as describe the scree plot (Retnawati, 2014). The output of factor analysis is the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalues and scree plot analysis results. If the cumulative percentage of the first factor eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first factor Eigenvalues is 20.220%. Because the Eigen value is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigen Values		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalues.

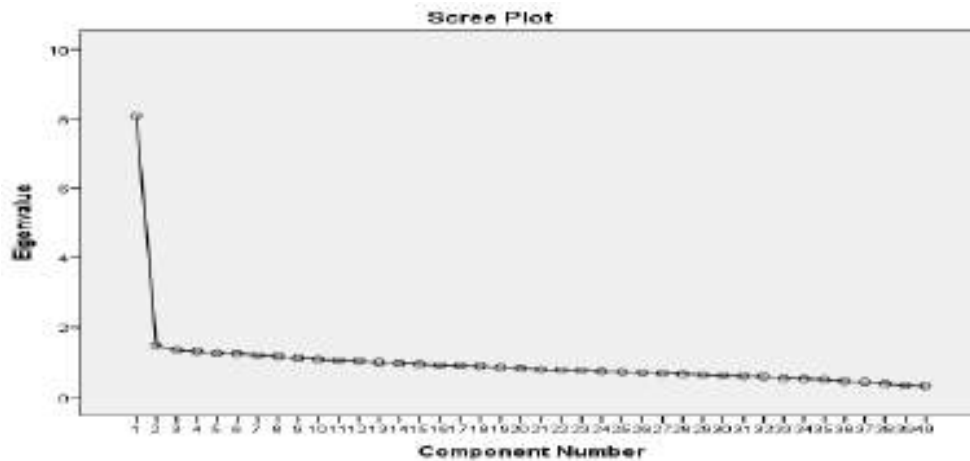


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the eigenvalues immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	<i>K1</i>	<i>K2</i>	<i>K3</i>	<i>K4</i>	<i>K5</i>	<i>K6</i>	<i>K7</i>	<i>K8</i>	<i>K9</i>	<i>K10</i>
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test is analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Corr is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items match the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	TOTAL MEASURE	MODEL S.E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	ENACT MATCH OBS%	EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.43	.07	.97	-.4	.98	-.5	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.3	.97	-.5	.43	.44	48.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	988	413	-.03	.07	.97	-.4	.98	-.4	.39	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.3	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.3	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	-.7	1.05	-.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.3	1.08	1.7	.56	.44	43.6	50.8	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	48.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.28	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.5	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.89	-8.3	.88	-8.2	.39	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level is analyzed using the Winsteps program, and the results obtained can be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-0.5	.97	-0.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-0.4	.98	-0.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-0.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-0.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-0.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-0.7	.96	-0.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-0.08	.07	.97	-0.4	.97	-0.4	.47	.44	50.8	50.9	Q20
21	1010	413	-0.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-0.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-0.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-0.4	.97	-0.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-0.01	.07	.97	-0.5	.97	-0.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-0.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-0.6	.96	-0.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-0.1	1.00	-0.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

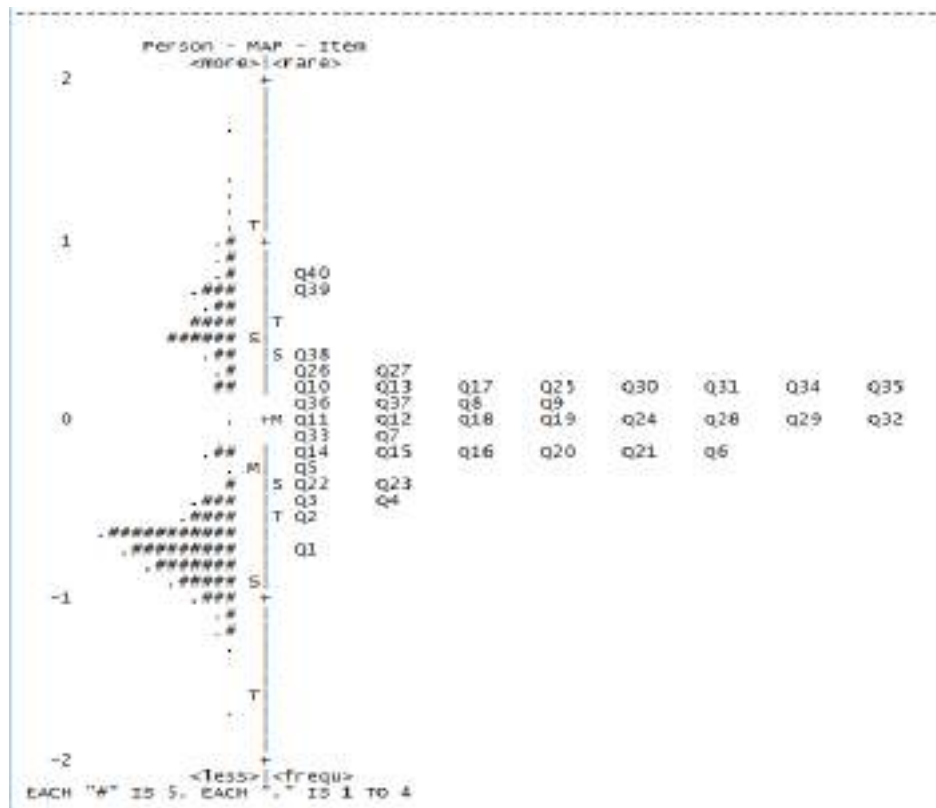


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure8 (column PT-Measure Corr). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than using classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen

that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability is measured by using a test that is expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

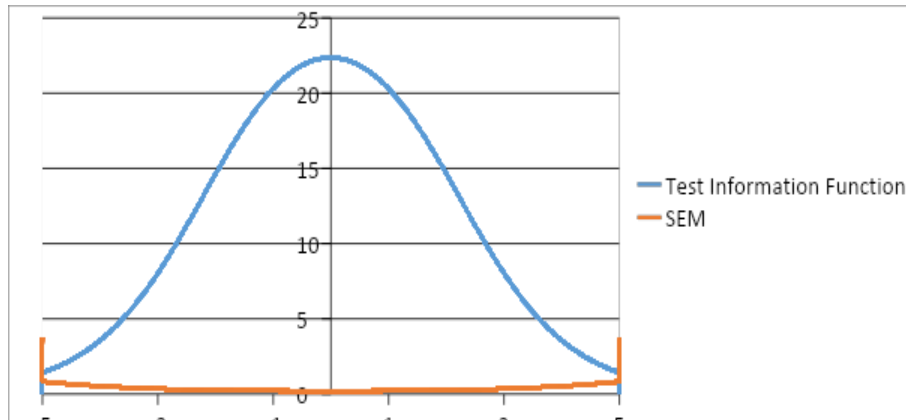


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function

expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in SMK is based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

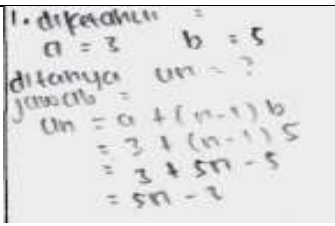
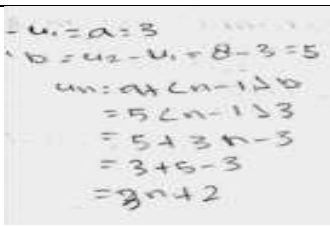
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	 <p>1. diketahui = $a = 3$ $b = 5$ ditanya $u_n = ?$ jawab = $u_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	 <p>- $u_1 = a = 3$ $b = u_2 - u_1 = 8 - 3 = 5$ $u_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>

Figure 10. Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only

writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

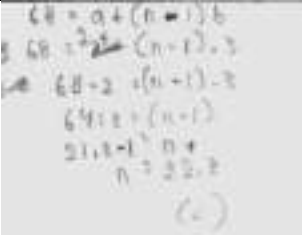
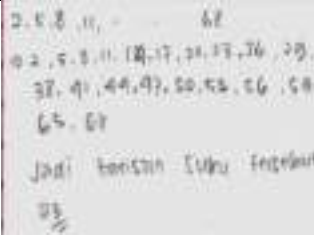
Question2:	Pattern 1	Pattern 2
<p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.</p> <p>The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>		

Figure 11. Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

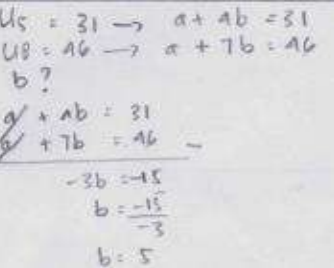
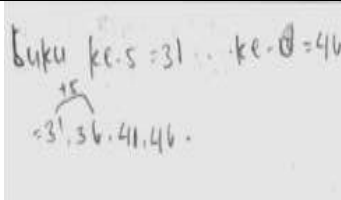
Question 3:	Pattern 1	Pattern 2
<p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>		

Figure 12. Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the n th term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the n th term.

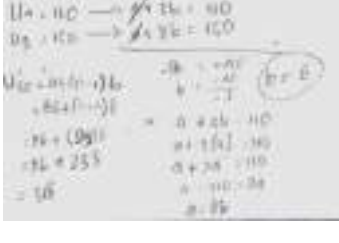
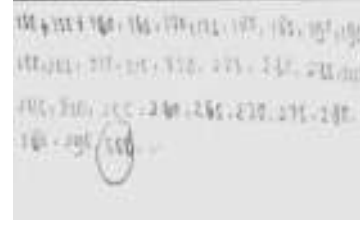
Question 4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>	 <p>Handwritten student work for Pattern 1. The student uses the general formula for the nth term of an arithmetic sequence, $U_n = a + (n-1)b$. They set up two equations: $U_4 = a + 3b = 110$ and $U_9 = a + 8b = 150$. They subtract the first equation from the second to get $5b = 40$, so $b = 8$. Then they substitute $b = 8$ back into the first equation to get $a + 24 = 110$, so $a = 86$. Finally, they calculate $U_{30} = 86 + (30-1) \cdot 8 = 86 + 232 = 318$.</p>	 <p>Handwritten student work for Pattern 2. The student lists terms of the sequence starting from the 4th term: 110, 118, 126, 134, 142, 150, 158, 166, 174, 182, 190, 198, 206, 214, 222, 230, 238, 246, 254, 262, 270, 278, 286, 294, 302, 310, 318. The 30th term, 318, is circled.</p>

Figure 13. Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

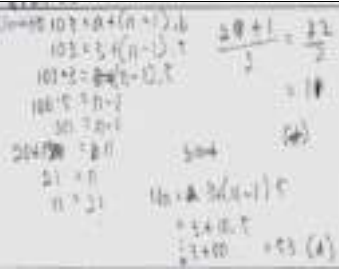

Question 5:	Answer Pattern 1	Answer Pattern 2
<p>An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...308</p> <p>A. 53 B. 52 C. 20 D. 11 E. 10</p> <p>Reason:</p>	 <p>Handwritten student work for Answer Pattern 1. The student uses the formula for the nth term, $U_n = a + (n-1)b$. They know $U_1 = 3$ and $U_n = 103$. They find the common difference $b = 5$. They then find the number of terms n by solving $103 = 3 + (n-1) \cdot 5$, which gives $n = 21$. The middle term is the 11th term, $U_{11} = 3 + (11-1) \cdot 5 = 53$.</p>	 <p>Handwritten student work for Answer Pattern 2. The student lists terms of the sequence: 3, 8, 13, 18, 23, 28, 33, 38, 43, 48, 53, 58, 63, 68, 73, 78, 83, 88, 93, 98, 103. The 11th term, 53, is circled.</p>

Figure 14. Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice tests are easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score, then the result is multiplied by 10 to get a value range of 0–10, or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8, 56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the

teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomus response test is more secondhand and complete than other assessment instruments, such as multiple choice tests (Gierl et al., 2017) or essay tests (Putri et al., 2020).

Discussion

This research is development research that aims to develop a good mathematical assessment instrument using polytomic responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). **That is, an item that is not in a good category with classical theory should be analyzed according to modern theory before revising or replacing the item.**

The study of assessment with an open response polytomus test was carried out by Yang et al. (2017). However, this Yang study has several fundamental differences, namely the research objectives and data analysis. The aim of this research is to diagnose the errors of university students in the concept of calculus, not to produce a good assessment instrument. The data analysis used parametric statistics (covariance), not item analysis (classic and modern). Because the objectives and data analysis are different, the results of this study cannot be compared with the results of any other study. However, this research has provided other

researchers with information in developing tests, such as the suitability of items with indicators, use of language, and preparation of answer choices.

There are other studies related to the response politomus test, namely Sarea (2018) and Saepuzaman (2021). Sarea's research states that the response polytomus test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomus test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas, and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 15 below.

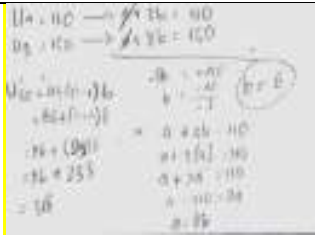
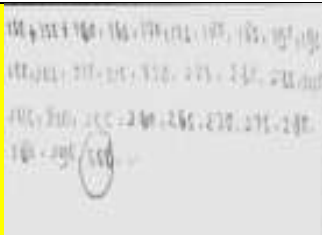
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>	 <p style="text-align: center;">(i)</p>	 <p style="text-align: center;">(ii)</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------

Figure 15. Student Answer Patterns with (i) Formulas, and (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, use a drawing or model, analogy, and formula. Syahlan (2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas

methods to solve difficult questions. If combined with student answers and Syahlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching a material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, conclusions are obtained, namely that (1) the open polytomous response test has a good category according to classical and modern theory. So, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices.

Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Andaria, M., & Hadiwinarto. (2020). Development of a two-tier multiple choice question assessment instrument to measure students' science process skills on acid-base material. *ISEJ: Indonesian Science Education Journal*, 1(3), 257-268. <https://siducat.org/index.php/isej/article/view/141>.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://doi.org/10.20956/jmsk.v9i2.3402>. This journal is not bilingual delete English title. doi link is not working
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *Jurnal Suluh Pendidikan/Journal of Educational Suluh*, 17(1), 32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: the cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. E. (2010). *Item response theory*. Oxford University Press. Add DOI link
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-Tier Diagnostic Test With Certainty of Response Index on The Concepts of Fluid. In C. Anwar (Ed.), *Young Scholar Symposium on Transdisciplinary in Education and Environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf1> Yellow highlighted in sentence case 2) green highlighted in sentence case and italics
- Dong, A., Jong, M. S., & King, R. B. (2020). How Does Prior Knowledge Influence Learning Engagement? The Mediating Roles of Cognitive Load and Help-Seeking. *Frontier Psychology Journal*. <https://doi.org/10.3389/fpsyg.2020.591203> Yellow highlighted in

sentence case. Also need vol.iss.pp.

- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Hong Kong: Assessment & Support Team. <https://wlts.edb.hkedcity.net/en/home/AandLI2.html>.
- Geofrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. *Researchgate Journal*.<https://doi.org/10.13140/RG.2.2.28470.22083>doi is not working. need vol.iss.pp.
- Gierl, M. J., Bulut,O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87 (6), 1082–1116. <https://doi.org/10.3102/0034654317726529>.
- Gonzalez, G. (2018). Understanding Teacher Noticing of Students' Prior Knowledge: Challenges and Possibilities.*The Mathematics Enthusiast*, 15 (3), 483-528. <https://doi.org/10.54870/1551-3440.1442>Yellow highlighted in sentence case
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12 (3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*.Kluwer.<https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*.Sage Publications.
- Jamhari, M. (2021). Developing the two-tier multiple choice tests in enhancing students' higher-order thinking skills on human digestive system. *Eduproxima: Jurnal Ilmiah Pendidikan IPA*, 3 (1), 50-64. <https://doi.org/10.29100/eduproxima.v3i1.1853>.This DOI link seems invalid. For works without DOIs from websites (not including databases), provide a URL in the reference (URL to full-text or abstract)
- Kartono. (2008). Equating the combined dichotomous and polytomuos item test model in an achievement test. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>.
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how?. In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485).Faculty of Mathematics and Natural Sciences. <http://seminar.uny.ac.id/icriems/proceeding2018>.
- Khusnah, M.(2019). The development of two-tiers diagnostic test for identifying tenth-grade

- student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://doi.org/10.31327/jme.v6i2.1607>. This DOI link seems invalid. For works without DOIs from websites (not including databases), provide a URL in the reference (URL to full-text or abstract)
- Linacre, J. M. (2012). *Winstep : Rasch-model computer programs*. Winsteps.Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset pemasaran*[Marketing research]. Eirlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically (Second edition)*. Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10 (2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. In Cari (Ed.), *Current research in Pandemic Covid-19 Era in Indonesia - 2nd ICOSETH 2020* (pp. 44-55). University of Sebelas Maret. <https://jurnal.uns.ac.id/ijssacs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Ozdemir, H., & Onder, N. (2017). Vocational high school students' perceptions of success in mathematics. *International Electronic Journal of Mathematics Education*, 12(3), 493-502. <https://doi.org/10.29333/iejme/627>.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah*[Content standards for primary and secondary education]. Indonesian Government publication service.
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>.
- Prasetya, A., Rosidin, U., & Herlina, K. (2019). Development of instrument assessment for learning the polytomous response models to train higher order thinking. In C. Anwar (Ed.), *Young Scholar Symposium on Transdisciplinary in Education and Environment (YSSTEE)* (pp. 1-11). UIN Raden Intan. http://repository.lppm.unila.ac.id/11982/1/33.%20Prasetya%202019_J.Phys.Conf.Ser..pdf
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch

- model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://doi.org/10.15294/jere.v9i2.46133>. This DOI link seems invalid. For works without DOIs from websites (not including databases), provide a URL in the reference (URL to full-text or abstract)
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6 (1), 97-107. <https://doi.org/10.15294/ujme.v6i1.12643>. This DOI link seems invalid. For works without DOIs from websites (not including databases), provide a URL in the reference (URL to full-text or abstract)
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst : Jurnal Pendidikan Fisika dan Terapannya/Karst : Journal of Physics Education and Its Application*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>.
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam questions for Islamic religious education and character at the elementary school level]. *An-Nahdhah/An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article/view/40>.
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Classical tes theory vs response theory? *Didaktika: Jurnal Kependidikan/Didactics: Journal of Education*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>. If a source is in another language, write the original title then add its English translation.
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic tests in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*,

4(6), 358-369.<http://idealmathedu.p4tkmatematika.org> A DOI is available for this reference. Use DOI link instead of URL link.

- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>.
- Treagust, D. (1988). Development and use of diagnostic tests to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep*[Analysis of research data with rash and winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122-1?inline=1>.
- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako/Electronic Journal of Tadulako Mathematics Education*, 6(4), 455-468. <http://jurnal.untad.ac.id/jurnal/index.php/JEPMT/article/view/14181>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian*[Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>.

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :
Class/Department :
School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

- | | |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>1. Given an arithmetic sequence: 3, 8, 13, 18,
The formula for the nth term of the sequence is
A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
C. $U_n = 4n - 1$</p> <p>Reason:</p> | <p>2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.
The number of terms in the sequence is...
A. 12 D. 23
B. 13 E. 24
C. 22</p> <p>Reason:</p> |
| <p>3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
A. 5 D. 8
B. 6 E. 11
C. 7</p> <p>Reason:</p> | <p>4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
A. 308 D. 344
B. 318 E. 354
C. 326</p> <p>Reason:</p> |
| <p>5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
Then the middle term of the sequence is ...
A. 53 D. 11
B. 52 E. 10
C. 20
D. 11
E. 10</p> <p>Reason:</p> | <p>6. Given the arithmetic sequence: 4, 10, 16, 22,
If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
A. 18 D. 24
B. 20 E. 26
C. 22</p> <p>Reason:</p> |
| <p>7. The nth term of an arithmetic series is $U_n = 3n - 5$.
The formula for the sum of the first n terms of the series is ...</p> | <p>8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....</p> |

- A. $S_n = \frac{n}{2}(3n - 7)$
 $\frac{n}{2}(3n - 3)$
 B. $S_n = \frac{n}{2}(3n - 5)$
 C. $S_n = \frac{n}{2}(3n - 4)$
 Reason:
- D. $S_n =$
 $\frac{n}{2}(3n - 3)$
 E. $S_n = \frac{n}{2}(3n - 3)$
 Reason:

- A. $5n - 20$
 B. $5n - 10$
 C. $2n - 30$
 D. $2n - 20$
 E. $2n - 10$

Reason:

9. The sum of all integers between 100 and 300 which are divisible by 5 is ...
 $S_n = \frac{n}{2}(3n - 7)$
 A. 8.200
 B. 8.000
 C. 7.800
 D. 7.600
 E. 7.400
 Reason:

10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
 A. 24
 B. 25
 C. 26
 D. 27
 E. 28
 Reason:

11. The middle term of an arithmetic sequence is 25. If the difference is 4 and the 5th term is 21. Then the sum of all the terms in the sequence is ...
 A. 175
 B. 189
 C. 275
 D. 295
 E. 375
 Reason:

12. A number of candies are distributed among five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces
 A. 60
 B. 65
 C. 70
 D. 75
 E. 80
 Reason:

13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...
 A. 564
 B. 276
 C. 48
 D. 45
 E. 36
 Reason:

14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...
 A. 9
 B. 10
 C. 11
 D. 12
 E. 13
 Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...
 A. 32
 B. 64
 C. 128
 D. 256
 E. 512
 Reason:

16. The value of the middle term of the geometric sequence: 6, 3, ..., 3/512 is ...
 A. $\frac{1}{16}$
 B. $\frac{2}{16}$
 C. $\frac{3}{16}$
 D. $\frac{4}{16}$
 E. $\frac{5}{16}$
 Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm
 A. 18
 B. 24
 C. 27,5
 D. 35
 E. 40,5
 Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...
 A. $\frac{3}{4}$
 B. $\frac{1}{4}$
 C. $\frac{1}{3}$
 D. $-\frac{1}{2}$
 E. $-\frac{3}{4}$
 Reason:

19. A ball falls from a height of 10 m and bounces back 3/4 times its previous height. The total number of paths until the ball stops is.... m
 A. 60
 B. 70
 C. 80
 D. 90
 E. 100

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.
 The value of $3a + b$ is ...
 A. 8
 B. 10
 C. 12
 D. 14
 E. 20

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L =$

$$\begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$$

If $K = L$, then c is ...

- A. 12
- B. 13
- C. 14
- D. 15
- E. 16

Reason:

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$.

Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$
- C. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$
- D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah

- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$
- B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$
- C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$
- D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
- E. $\begin{bmatrix} 0 & 7 & 6 \\ 1 & 3 & 1 \\ 2 & 4 & 5 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah

- A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$
- B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$
- C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$
- D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
- E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...

- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$
- B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$
- C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$
- D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
- E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix

- A. 0
- B. 1
- C. 2
- D. 2
- E. 4

Reason:

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and

$$C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$$

If $A + B = C$, then $x + y = \dots$

- A. -5
- B. -1
- C. 1
- D. 3
- E. 5

Reason:

26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$

then $A(B - C) = \dots$

- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$
- B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$
- C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$
- D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
- E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix.

Value of x that satisfies is

- A. -5
- B. -4
- C. -3
- D. 3
- E. 4

Reason:

30. Transpose matrix P adalah P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ...

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$
- B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$
- C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$
- D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
- E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$.
Inverse matrix AB adalah $(AB)^{-1} = \dots$
- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$
- Reason:

32. The roots of the quadratic equation $3x^2 - 4x - 4 = 0$ are ...
- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$
- Reason:

33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...
- A. -2 dan $\frac{5}{6}$ D. -2 dan $-\frac{6}{5}$
 B. 2 dan $-\frac{5}{6}$ E. -2 dan $\frac{6}{5}$
 C. 2 dan $\frac{6}{5}$
- Reason:

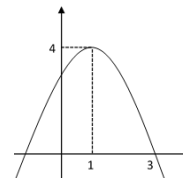
34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...
- A. -4 D. 2
 B. -2 E. 4
 C. 0
- Reason:

35. The roots of the quadratic equation $2x^2 - 3x - 9 = 0$ are x_1 dan x_2 . Value of $x_1^2 + x_2^2$ is ...
- A. $11\frac{1}{4}$ D. $-6\frac{3}{4}$
 B. $6\frac{3}{4}$ E. $-11\frac{1}{4}$
 C. $2\frac{1}{4}$
- Reason:

36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α dan β . The quadratic equation $(\alpha - 2)$ dan $(\beta - 2)$ is ...
- a. $x^2 + 6x + 5 = 0$
 b. $x^2 + 6x + 7 = 0$
 c. $x^2 + 6x + 11 = 0$
 d. $x^2 - 2x + 3 = 0$
 e. $x^2 + 2x + 11 = 0$
- Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...
- A. $y = x^2 - 2x + 1$
 B. $y = x^2 - 2x + 3$
 C. $y = x^2 + 2x - 1$
 D. $y = x^2 + 2x + 1$
 E. $y = x^2 + 2x + 3$
- Reason:

38. The figure below is a graph of the quadratic equation ? ...
- A. $y = x^2 + 2x + 3$
 B. $y = x^2 - 2x - 3$
 C. $y = -x^2 + 2x - 3$
 D. $y = -x^2 - 2x + 3$
 E. $y = -x^2 + 2x + 3$
- Reason:



39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0, -4)$ then the value of $f(7)$ is ...
- A. -16 D. -19
 B. -17 E. -20
 C. -18
- Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...
- A. $y = -x^2 + 2x - 3$ D. $y = -x^2 - 2x - 5$
 B. $y = -x^2 + 2x + 3$ E. $y = -x^2 - 2x + 5$
 C. $y = -x^2 - 2x + 3$
- Reason:

The Development of an Assessment Instrument Using Polytomous Response in Mathematics

Abstract: This research is a development research aimed at developing a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. This research design uses the Plomp model, which consists of five stages, namely: (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data was collected through questionnaire and test. The questionnaire was to identify instruments commonly used by teachers and to validate them by experts of mathematics and educational evaluation. The test used the open polytomous response test as many as 40 items. The data was analyzed using both classical and modern theories. The results show that the open polytomous response test has a good category according to classical and modern theory, it can provide information on the actual competence of students; this is observed from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: assessment instrument, classical and modern theory, vocational school, the polytomous responses

Introduction

Assessment is an important activity and needs to be done by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools is divided into three types: assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). These three types of assessments aim to

provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013), as shown in the following assessment pyramid (Figure 1).



Figure 1. Assessement Pyramid

Assessment can be used with test. A test is a tool or procedure used to find out or measure students' abilities in certain areas with certain rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the potions is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has strengths or weaknesses compared to each other. The strength of multiple-choice test over essay is that multiple-choice test can be conducted for many students, are more objective, and the test results can be known more quickly. Unfortunately they have a weakness, namely that the multiple-choice test is not able to see the actual abilities of students and the answers tend to be guessed or tried out (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (getting a score of 1 for the correct answer, and a score of 0 for the wrong answer choice). Multiple-choice test with only two answer choices are called "dichotomous test," and multiple-choice test with more than two answer choices are called "polytomous test" (Kartono, 2008).

Until now, multiple-choice test have been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons

and are called the polytomous response test (Suwanto, 2012). The polytomous response test score is 1 - 4. Score of 4 for the correct answer and reason, score of 3 for the correct answer but the wrong reason, score of 2 for the wrong answer but the correct reason, and score of 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as, test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be known in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test to be the open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). **The studies on the open polytomous response test that have been carried**

out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed the open polytomous response test for students in college, and senior or junior school. Students in college, and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government wants vocational schools to not be left behind in academic subjects such as mathematics. The government's desire is to improve the way of assessing student learning in vocational school, and so far the assessment method used is the polytomus test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmawati, 2020). This reason causes students to tend to answer the test by guessing. Therefore, to avoid these student tendencies, it is necessary to develop a polytomus response test (closed or open). Taking into account the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing the open polytomous response test for students in vocational schools.

The test instrument developed must be accountable as a good test, and be necessary to analyze the quality of the item (Rosidin, 2017). There are two theories for analyzing item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item

discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities and is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

The aim of this development research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows: (1) does the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory?, and (2) does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools ?

Methodology

Research Design

This research is a research and development that refers to Plomp's model (Plomp, 2013), with the research procedure consisting of five stages, namely preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

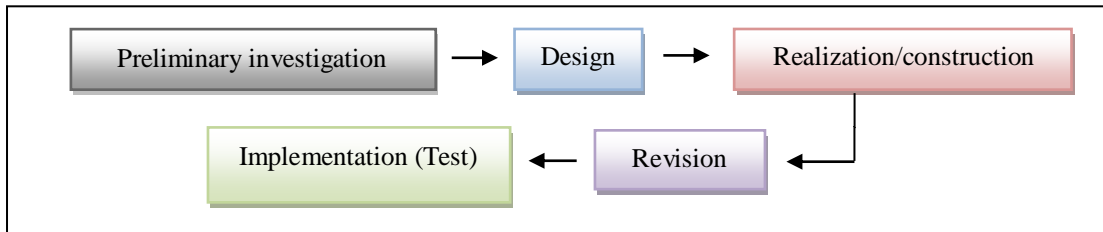


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the assessment instruments used by teachers so far. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and analyze the results of the test.

Research Subject

The subjects of the study were students at a vocational school in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data was collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two people who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining content validity, the instrument was tested on students. Then, it was continued by determining the validity of the construct and its reliability. It was aimed to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and

functions, and matrices. Each item contained five answer choices along with the reasons. Student scores referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected research data was analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. *Questionnaire data analysis (qualitative analysis)*

There are two sets of questionnaire data, namely: identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. *Test data analysis (empirical analysis)*

After the content validity test, it was followed by the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to have a good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program was used for classical theory and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 *Analysis of test data with classical theory*

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.

- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional

solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that so far, the teacher had never used the polytomous response. As many as 80% of teachers used essay test and 20% of teachers used multiple-choice test, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as: teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

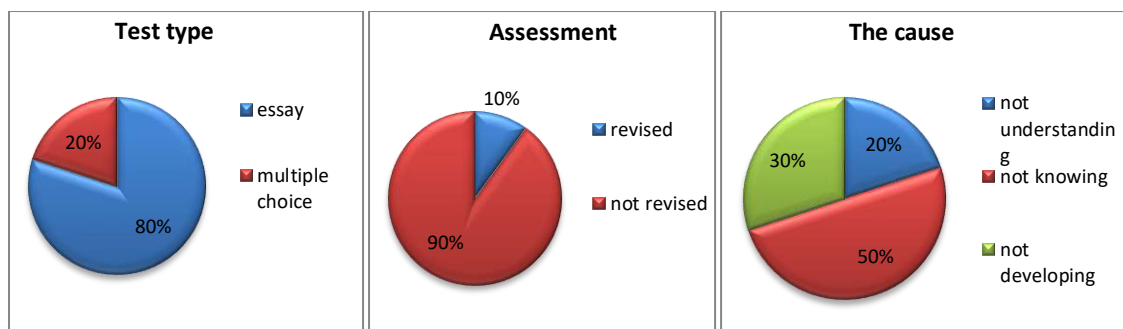


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's Alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised

4	0.540	good	0.304	Good	24	0.438	good	-0.071	revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are included in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

Analysis of Test Data with Modern Theory

The Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first factor Eigenvalue is greater than

20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

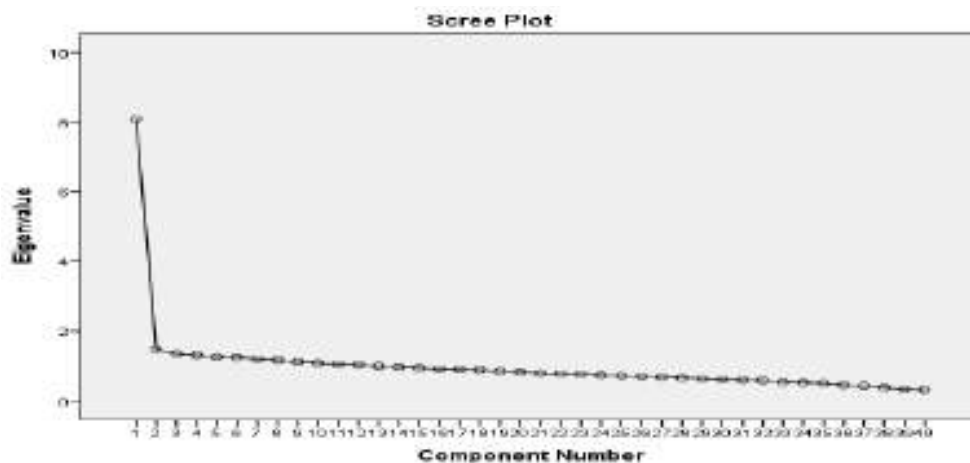


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be

determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	<i>K1</i>	<i>K2</i>	<i>K3</i>	<i>K4</i>	<i>K5</i>	<i>K6</i>	<i>K7</i>	<i>K8</i>	<i>K9</i>	<i>K10</i>
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Corr is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MOOGL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH OBS%	EXACT MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.43	.07	.97	-1.4	.98	-1.5	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.8	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.3	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.3	.97	-.5	.45	.44	48.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	964	413	-.01	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.3	1.10	1.6	.49	.44	42.8	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.3	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.0	50.9	Q26
27	936	413	.20	.07	1.02	-.3	1.02	-.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	849	413	.14	.07	1.06	.9	1.01	.9	.30	.44	45.5	50.9	Q30
31	958	413	.11	.07	1.07	1.1	1.07	1.1	.25	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.04	.07	1.04	-.6	1.04	-.6	.41	.44	48.2	51.0	Q33
34	959	413	-.09	.07	1.06	1.0	1.07	1.1	.35	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.69	-6.3	.66	-6.2	.39	.43	38.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained was be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

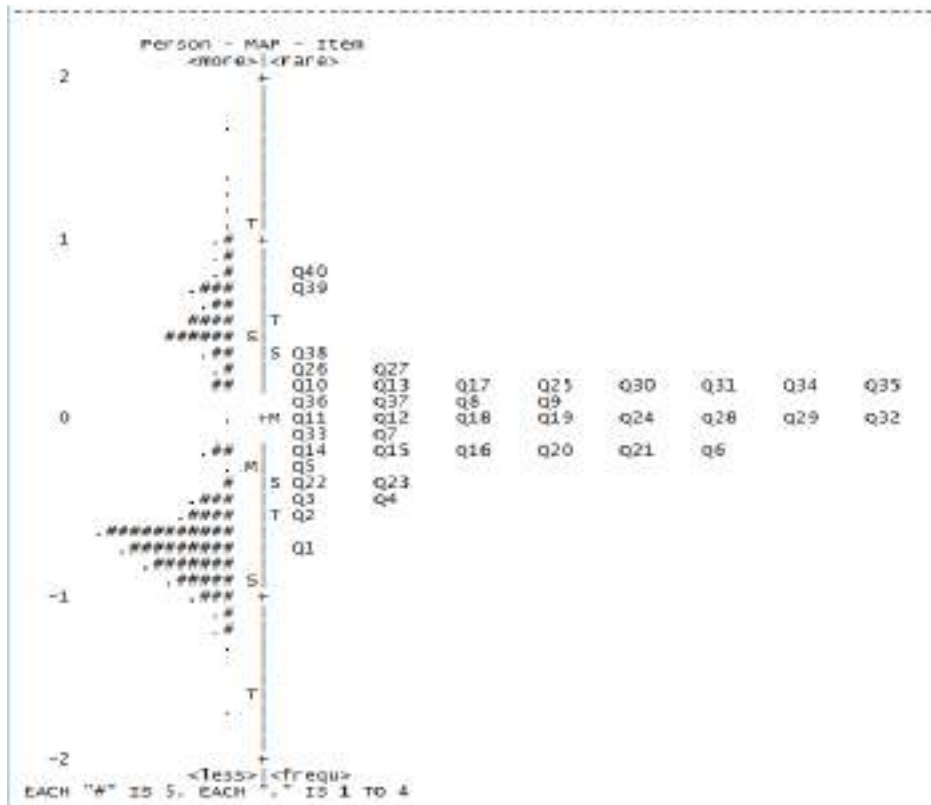


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (column PT-Measure Corr). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than using classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

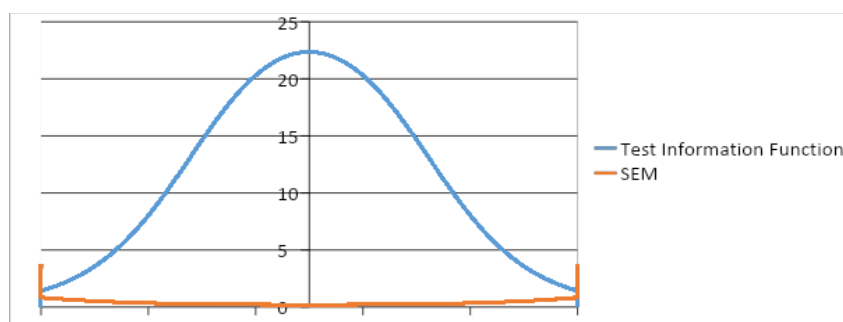


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in vocational school based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

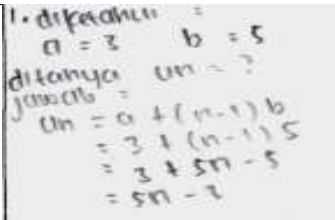
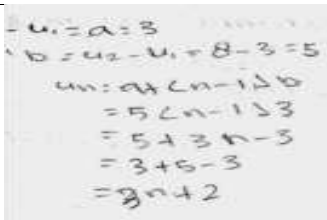
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, ... The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p>	 <p>1. diketahui : $a = 3$ $b = 5$ ditanya $U_n = ?$ jawab : $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	 <p>$U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>
Reason:		

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

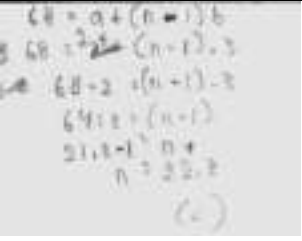
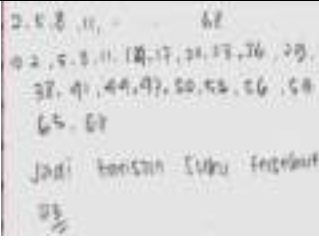
Question 2:	Pattern 1	Pattern 2
<p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>		

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

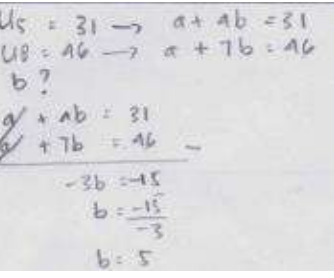
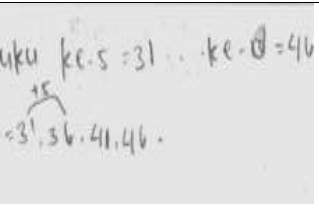
Question 3:	Pattern 1	Pattern 2
<p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>		

Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the n th term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the n th term.

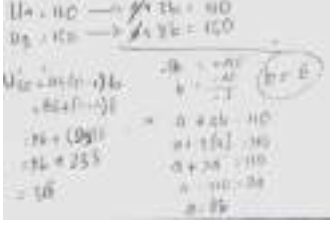
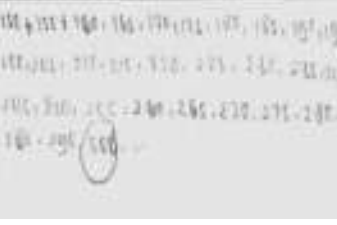
Question 4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>	 <p>Handwritten student work for Pattern 1. It shows the general formula for an arithmetic sequence: $U_n = a + (n-1)b$. It then sets up two equations: $110 = a + 3b$ and $150 = a + 5b$. By subtracting the first equation from the second, it finds $40 = 2b$, so $b = 20$. Substituting $b = 20$ into the first equation gives $110 = a + 60$, so $a = 50$. Finally, it calculates the 30th term: $U_{30} = 50 + (30-1) \cdot 20 = 50 + 580 = 630$.</p>	 <p>Handwritten student work for Pattern 2. It lists terms of the sequence: 110, 130, 150, 170, 190, 210, 230, 250, 270, 290, 310, 330, 350, 370, 390, 410, 430, 450, 470, 490, 510, 530, 550, 570, 590, 610, 630. The 30th term, 630, is circled.</p>

Figure 13. An Example of Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

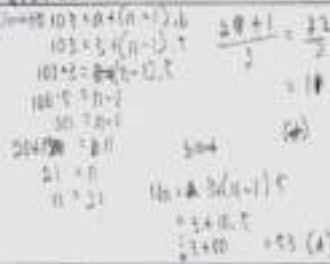

Question 5:	Answer Pattern 1	Answer Pattern 2
<p>An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...308</p> <p>A. 53 B. 52 C. 20 D. 11 E. 10</p> <p>Reason:</p>	 <p>Handwritten student work for Answer Pattern 1. It uses the formula for the nth term: $U_n = a + (n-1)b$. It sets up the equation $103 = 3 + (n-1) \cdot 5$. Solving for n, it gets $100 = 5(n-1)$, $20 = n-1$, so $n = 21$. The middle term is the 11th term: $U_{11} = 3 + (11-1) \cdot 5 = 3 + 50 = 53$.</p>	 <p>Handwritten student work for Answer Pattern 2. It lists terms of the sequence: 3, 8, 13, 18, 23, 28, 33, 38, 43, 48, 53, 58, 63, 68, 73, 78, 83, 88, 93, 98, 103. The middle term, 53, is circled.</p>

Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice test are easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score, then the result is multiplied by 10 to get a value range of 0–10, or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8, 56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the

teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomus response test is more secondhand and complete than other assessment instruments, such as multiple choice test (Gierl et al., 2017) or essay test (Putri et al., 2020).

Discussion

This research is a development research aimed at developing a good mathematical assessment instrument using polytomic responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to modern theory before revising or replacing the item.

Research on learning assessment with open response politomus was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test, namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

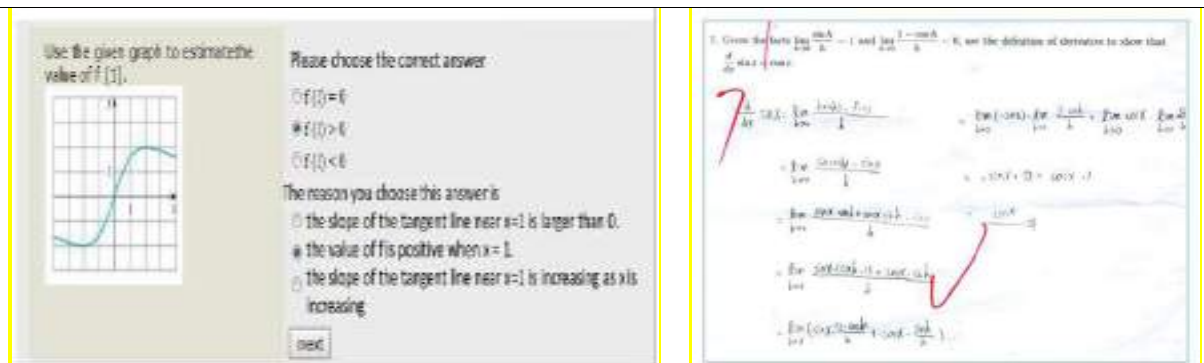


Figure 15. An Example of Student Answers on The Two-Tier Test and

Another study on learning assessment with open response polytomus was conducted by Ayanwale (2021). Ayanwale's research compares two polytomous response test analysis methods, namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

There are other studies related to classical and modern theory, namely Sarea (2018) and Saepuzaman et al. (2021). Sarea's research states that the response polytomus test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomus test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas, and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and

students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

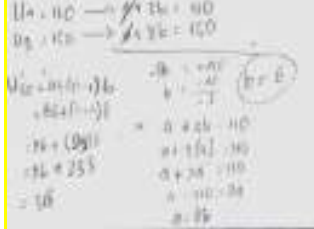
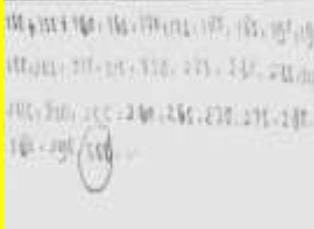
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>	 <p style="text-align: center;">(i)</p>	 <p style="text-align: center;">(ii)</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, use a drawing or model, analogy, and formula. Syahlan (2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syahlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching a material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The

description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, conclusions are obtained: (1) the open polytomous response test has a good category according to classical and modern theory. So, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://journal.unhas.ac.id/index.php/jmsk/article/view/3402/1939>
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *Jurnal Suluh Pendidikan*, 17(1), 32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://ejournal.upsi.edu.my/index.php/EJSMT/article/view/5029>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: the cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>

- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Hong Kong: Assessment & Support Team. [https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_\(1\)%20Lorna.pdf](https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_(1)%20Lorna.pdf)
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geofrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. <https://www.researchgate.net/publication/351451835>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082–1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-

- tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1177/014662168200600401>
- Ikmalwati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on mathematics learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Equating the combined dichotomous and polytomous item test model in an achievement test. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>.
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how?. In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. http://seminar.uny.ac.id/icriems/sites/seminar.uny.ac.id/icriems/files/proceeding2018/M_E69_Asma%20Khiyarunnisa.pdf
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://usnsj.com/index.php/JME/article/view/1607>
- Linacre, J. M. (2012). *Winstep : Rasch-model computer programs*. Winsteps. Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Eirlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically (Second edition)*. Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. In Cari (Ed.), *Current research in Pandemic Covid-19 Era in Indonesia - 2nd ICOSETH 2020* (pp. 44-55). University of Sebelas Maret. <https://jurnal.uns.ac.id/ijscs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.

- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Content standards for primary and secondary education]. Indonesian Government publication service. https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud_Tahun2016_Nomor021_Lampiran.pdf
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://journal.unnes.ac.id/sju/index.php/jere/article/view/46133>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://journal.unnes.ac.id/sju/index.php/ujme/article/view/12643>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163. <http://dx.doi.org/10.30587/postulat.v1i2.2094>
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst : Jurnal Pendidikan Fisika dan Terapannya/ Karst: Journal of Physics Education and Its Application*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article%20view/40>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: teori tes klasik and respon [Characteristics of items: classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3),

1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>

- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <http://dx.doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4 (6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9 (4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep* [Analysis of research data with rash and winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122-1?inline=1>
- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <http://jurnal.untad.ac.id/jurnal/index.php/JEPMT/article/view/14181>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :
Class/Department :
School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

1. Given an arithmetic sequence: 3, 8, 13, 18,
The formula for the nth term of the sequence is
A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
C. $U_n = 4n - 1$
Reason:
2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.
The number of terms in the sequence is...
A. 12 D. 23
B. 13 E. 24
C. 22
Reason:
3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
A. 5 D. 8
B. 6 E. 11
C. 7
Reason:
4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
A. 308 D. 344
B. 318 E. 354
C. 326
Reason:
5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
Then the middle term of the sequence is ...
A. 53 D. 11
B. 52 E. 10
C. 20
D. 11
E. 10
Reason:
6. Given the arithmetic sequence: 4, 10, 16, 22,
If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
A. 18 D. 24
B. 20 E. 26
C. 22
Reason:
7. The nth term of an arithmetic series is $U_n = 3n - 5$.
The formula for the sum of the first n terms of the series is ...
A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
C. $S_n = \frac{n}{2}(3n - 4)$
Reason:
8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....
A. $5n - 20$ D. $2n - 20$
B. $5n - 10$ E. $2n - 10$
C. $2n - 30$
Reason:
9. The sum of all integers between 100 and 300 which are divisible by 5 is ... $S_n = \frac{n}{2}(3n - 7)$
A. 8.200 D. 7.600
B. 8.000 E. 7.400
C. 7.800
Reason:
10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
A. 24 D. 27
B. 25 E. 28
C. 26
Reason:

11. The middle term of an arithmetic sequence is 25.
If the difference is 4 and the 5th term is 21.
Then the sum of all the terms in the sequence is ...
A. 175 D. 295
B. 189 E. 375
C. 275
Reason:
12. A number of candies are distributed among five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces
A. 60 D. 75
B. 65 E. 80
C. 70
Reason:
13. The sum of the first n terms of a series is $2n^2-n$.
So the 12th term of the series is...
A. 564 D. 45
B. 276 E. 36
C. 48
Reason:
14. The number of terms in the geometric sequence:
3, 6, 12, ..., 3072 is ...
A. 9 D. 12
B. 10 E. 13
C. 11
Reason:
15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...
A. 32 D. 256
B. 64 E. 512
C. 128
Reason:
16. The value of the middle term of the geometric sequence:
6, 3, ..., $3/512$ is ...
A. $\frac{1}{16}$ D. $\frac{4}{16}$
B. $\frac{2}{16}$ E. $\frac{5}{16}$
C. $\frac{3}{16}$
Reason:
17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm
A. 18 D. 35
B. 24 E. 40,5
C. 27,5
Reason:
18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...
A. $\frac{3}{4}$ D. $-\frac{1}{4}$
B. $\frac{1}{4}$ E. $-\frac{3}{4}$
C. $\frac{1}{3}$
Reason:
19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is... m
A. 60 D. 90
B. 70 E. 100
C. 80
Reason:
20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.
The value of $3a + b$ is ...
A. 8 D. 14
B. 10 E. 20
C. 12
Reason:
21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$.
If $K = L$, then c is ...
A. 12 D. 15
B. 13 E. 16
C. 14
Reason:
22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$.
Then $(A + C) - (A + B)$ is ...
A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$
C. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$
Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah
- ...
- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$
- Reason:
24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.
- If $A + B = C$, then $x + y = \dots$
- A. -5 D. 3
 B. -1 E. 5
 C. 1
- Reason:
25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah
- ...
- A. $\begin{bmatrix} 13 & 42 \\ 26 & 84 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \\ 30 & 36 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \\ 26 & 42 \end{bmatrix}$ E. $\begin{bmatrix} 13 & 84 \\ 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \\ 26 & 42 \end{bmatrix}$
- Reason:
26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$ then $A(B - C) = \dots$
- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$
- Reason:
27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...
- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$
- Reason:
28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix. Value of x that satisfies is ...
- A. -5 D. 3
 B. -4 E. 4
 C. -3
- Reason:
29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ...
- A. 0 D. 2
 B. 1 E. 4
 C. 2
- Reason:
30. Transpose matrix P adalah P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ...
- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$
- Reason:
31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$. Inverse matrix AB adalah $(AB)^{-1} = \dots$
- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$
- Reason:
32. The roots of the quadratic equation $3x^2 - 4x - 4 = 0$ are ...
- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$
- Reason:
33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...
- A. -2 dan $\frac{5}{6}$ D. -2 dan $-\frac{6}{5}$
 B. 2 dan $-\frac{5}{6}$ E. -2 dan $\frac{6}{5}$
 C. 2 dan $\frac{6}{5}$
- Reason:
34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...
- A. -4 D. 2
 B. -2 E. 4
 C. 0
- Reason:

35. The roots of the quadratic equation $2x^2 - 3x - 9 = 0$ are x_1 dan x_2 . Value of $x_1^2 + x_2^2$ is ...

- A. $11\frac{1}{4}$
- B. $6\frac{3}{4}$
- C. $2\frac{1}{4}$
- D. $-6\frac{3}{4}$
- E. $-11\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...

- A. $y = x^2 - 2x + 1$
- B. $y = x^2 - 2x + 3$
- C. $y = x^2 + 2x - 1$
- D. $y = x^2 + 2x + 1$
- E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0, -4)$ then the value of $f(7)$ is ...

- A. -16
- B. -17
- C. -18
- D. -19
- E. -20

Reason:

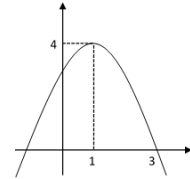
36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α dan β . The quadratic equation $(\alpha - 2)$ dan $(\beta - 2)$ is ...

- a. $x^2 + 6x + 5 = 0$
- b. $x^2 + 6x + 7 = 0$
- c. $x^2 + 6x + 11 = 0$
- d. $x^2 - 2x + 3 = 0$
- e. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation ? ...

- A. $y = x^2 + 2x + 3$
- B. $y = x^2 - 2x - 3$
- C. $y = -x^2 + 2x - 3$
- D. $y = -x^2 - 2x + 3$
- E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...

- A. $y = -x^2 + 2x - 3$
- B. $y = -x^2 + 2x + 3$
- C. $y = -x^2 - 2x + 3$
- D. $y = -x^2 - 2x - 5$
- E. $y = -x^2 - 2x + 5$

Reason:



SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>

5th round corrections request for the manuscript ID# 21112502244011

2 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>
Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Thu, Apr 14, 2022 at 5:36 PM

Dear Dr. Sutiarso,

Please see the attached file as the 5th round corrections.

According to last revision of our reviewer, the paper needs a proofreading by a language expert. Please send the certificate of the proofreading with your revised paper.

Please remove the old highlights and re-highlight for new edited parts. We don't need a new correction report.

We are looking forward to getting your second revised paper until **April 20, 2022**.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com

On 4/11/2022 5:35 PM, SUGENG SUTIARSO wrote:

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.

I have revised the article according to the reviewer's suggestion. Here I attach 4th round of article revisions.

Best regards,

Sugeng Sutiarso
Lampung University

On Mon, Apr 4, 2022 at 3:49 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarso,

Please see the attached file as the second round corrections.

Please remove the old highlights and re-highlight for new edited parts. We don't need a new correction report.

We are looking forward to getting your second revised paper until **April 11, 2022**.**PS. If the all corrections can't be done, the editorial process will be cancelled.**

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com

On 29-Mar-22 12:25 PM, SUGENG SUTIARSO wrote:

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.

I have revised the article according to the reviewer's suggestion. Here I attach (1) 3rd round of article revisions,

and (2) 3rd correction report.

Best regards,

Sugeng Sutiarmo
Lampung University

On Wed, Mar 16, 2022 at 7:00 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarmo,

Please see the attached file as the second round corrections.

Please remove the old highlights and re-highlight for new edited parts. We need a new correction report.

We are looking forward to getting your second revised paper until **March 30, 2022**.

PS. If the all corrections can't be done, the editorial process will be cancelled.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com

On 15-Mar-22 3:14 PM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research.

I apologize for my error in citation (there is a problem in my computer).
Here I re-send my article.
Thank you for this opportunity to improve.

Best regards,
Sugeng Sutiarmo
Lampung University, Indonesia.

On Tue, Mar 15, 2022 at 1:57 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

In-text citations are not visible in the edited file (we guess it's because of the program you were using). Could you correct the citations and re-send it please urgently?

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 13-Mar-22 3:51 PM, SUGENG SUTIARSO wrote:

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.

I have revised the second round of corrections according to the reviewer's suggestion.
Here I attach a correction report and revised article.

Thank you.

Best regards,

Sugeng Sutiarmo
Lampung University

On Mon, Feb 28, 2022 at 7:59 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarmo,

Please see the attached file as the second round corrections.

Please remove the old highlights and re-highlight for new edited parts. We need a new correction report.

We are looking forward to getting your second revised paper until **March 14, 2022**.

Best regards,

Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com

On 22-Feb-22 4:48 PM, SUGENG SUTIARSO wrote:

Dear Ahmet Savas, Ph.D.
Editor, European Journal of Educational Research.

I have revised the article according to the reviewer's suggestion. Here I attach (1) a revised article, (2) a correction report, and (3) a proofreading certificate from my university's language center.

Best regards,

Sugeng Sutiarmo
Lampung University

 **5TH ROUND_EU-JER_21112502244011.docx**
1703K

SUGENG SUTIARSO <sugeng.sutiarmo@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Wed, Apr 20, 2022 at 2:31 PM

Dear Ahmet Savas, Ph.D.

Editor, European Journal of Educational Research.

I have revised the article according to the reviewer's suggestion.

I attach: (1) 5th of article revisions, and (2) a proofreading certificate from my university's language center.

Best regards,

The Development of an Assessment Instrument Using Polytomous Response in Mathematics

Abstract: This research is a development research aimed at developing a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. This research design uses the Plomp model, which consists of five stages, namely: (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data was collected through questionnaire and test. The questionnaire was to identify instruments commonly used by teachers and to validate them by experts of mathematics and educational evaluation. The test used the open polytomous response test as many as 40 items. The data was analyzed using both classical and modern theories. The results show that the open polytomous response test has a good category according to classical and modern theory, it can provide information on the actual competence of students; this is observed from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses

Introduction

Assessment is an important activity and needs to be done by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools is divided into three types: assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). These three types of assessments aim to

provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013), as shown in the following assessment pyramid (Figure 1).



Figure 1. Assesment Pyramid

Assessment can be used with test. A test is a tool or procedure used to find out or measure students' abilities in certain areas with certain rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the potions is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has strengths or weaknesses compared to each other. The strength of multiple-choice test over essay is that multiple-choice test can be conducted for many students, are more objective, and the test results can be known more quickly. Unfortunately they have a weakness, namely that the multiple-choice test is not able to see the actual abilities of students and the answers tend to be guessed or tried out (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (getting a score of 1 for the correct answer, and a score of 0 for the wrong answer choice). Multiple-choice test with only two answer choices are called "dichotomous test," and multiple-choice test with more than two answer choices are called "polytomous test" (Kartono, 2008).

Until now, multiple-choice test have been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons

and are called the polytomous response test (Suwanto, 2012). The polytomous response test score is 1 - 4. Score of 4 for the correct answer and reason, score of 3 for the correct answer but the wrong reason, score of 2 for the wrong answer but the correct reason, and score of 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as, test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be known in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test to be the open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics

material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed the open polytomous response test for students in college, and senior or junior school. Students in college, and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government wants vocational schools to not be left behind in academic subjects such as mathematics. The government's desire is to improve the way of assessing student learning in vocational school, and so far the assessment method used is the polytomus test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmawati, 2020). This reason causes students to tend to answer the test by guessing. Therefore, to avoid these student tendencies, it is necessary to develop a polytomus response test (closed or open). Taking into account the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing the open polytomous response test for students in vocational schools.

The test instrument developed must be accountable as a good test, and be necessary to analyze the quality of the item (Rosidin, 2017). There are two theories for analyzing item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a

measurement theory to assess students' abilities by comparing students' abilities with their group abilities and is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

The aim of this development research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows:(1) does the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory?, and (2) does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools ?

Methodology

Research Design

This research is a research and development that refers to Plomp's model (Plomp, 2013), with the research procedure consisting of five stages, namely preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

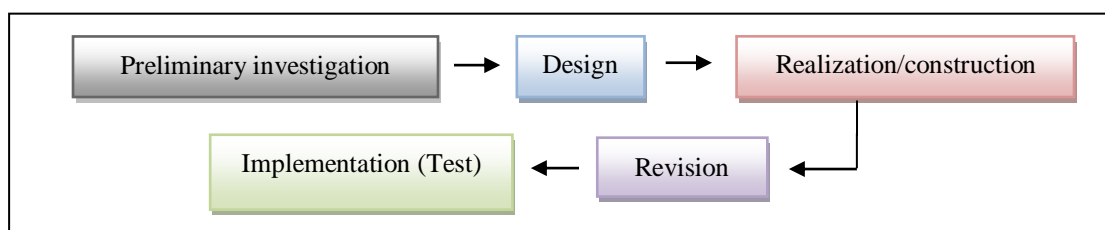


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the assessment instruments used by teachers so far. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and analyze the results of the test.

Research Subject

The subjects of the study were students at a vocational school in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data was collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two people who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining content validity, the instrument was tested on students. Then, it was continued by determining the validity of the construct and its reliability. It was aimed to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons.

Student scores referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected research data was analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. *Questionnaire data analysis (qualitative analysis)*

There are two sets of questionnaire data, namely: identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item.

Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. *Test data analysis (empirical analysis)*

After the content validity test, it was followed by the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to have a good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60(Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the IteMan program was used for classical theory and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 *Analysis of test data with classical theory*

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3,

and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that so far, the teacher had never used the polytomous response. As many as 80% of teachers used essay test and 20% of teachers used multiple-choice test, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as: teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

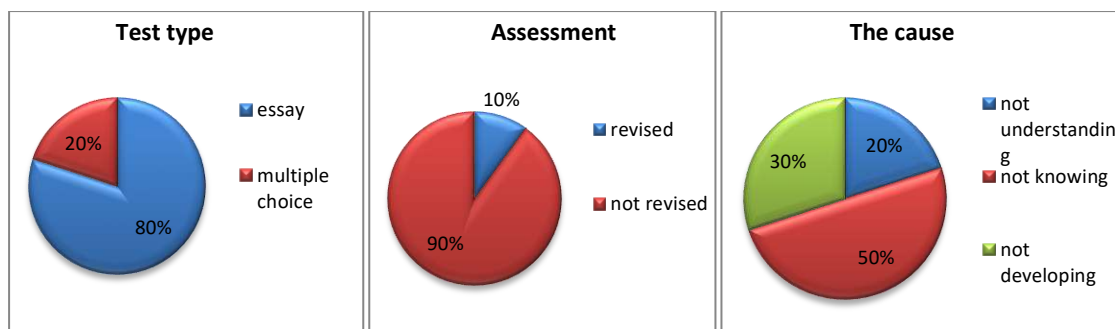


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's Alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised

4	0.540	good	0.304	Good	24	0.438	good	-0.071	revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are included in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

Analysis of Test Data with Modern Theory

The Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first factor Eigenvalue is greater than

20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

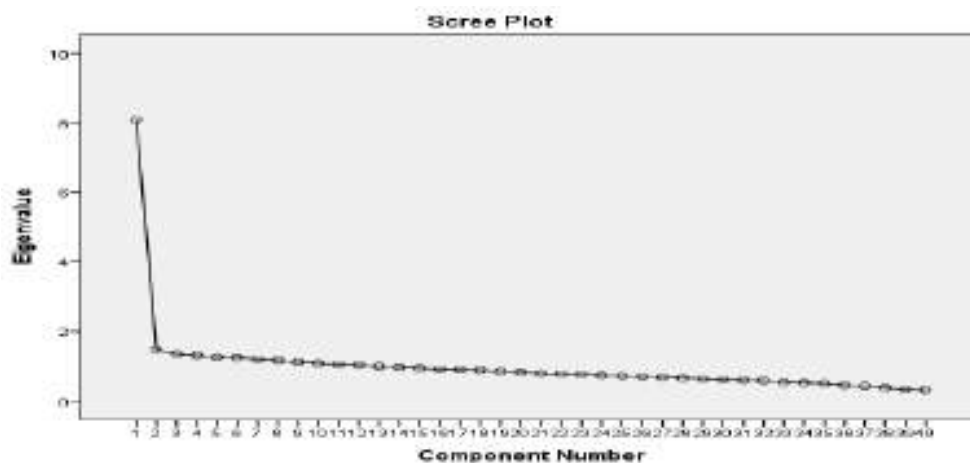


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be

determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	<i>K1</i>	<i>K2</i>	<i>K3</i>	<i>K4</i>	<i>K5</i>	<i>K6</i>	<i>K7</i>	<i>K8</i>	<i>K9</i>	<i>K10</i>
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Corr is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MOOGL S. E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT MATCH OBS%	EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.43	.07	.97	-2.4	.98	-3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-3.0	.94	-3.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.8	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.3	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.9	.97	-.5	.45	.44	48.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	964	413	-.01	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.3	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.3	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.3	1.08	1.2	.56	.44	43.6	50.6	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	-.16	.07	.97	-.4	.97	-.4	.31	.44	50.0	50.9	Q26
27	936	413	-.20	.07	1.02	-.3	1.02	-.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	-.14	.07	1.06	.9	1.01	.9	.30	.44	45.5	50.9	Q30
31	958	413	-.11	.07	1.07	1.3	1.07	1.1	.35	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.04	.07	1.04	-.6	1.04	-.6	.41	.44	48.2	51.0	Q33
34	959	413	-.09	.07	1.06	1.0	1.07	1.1	.35	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.5	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.69	-6.3	.66	-6.2	.39	.43	38.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.13	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained was be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

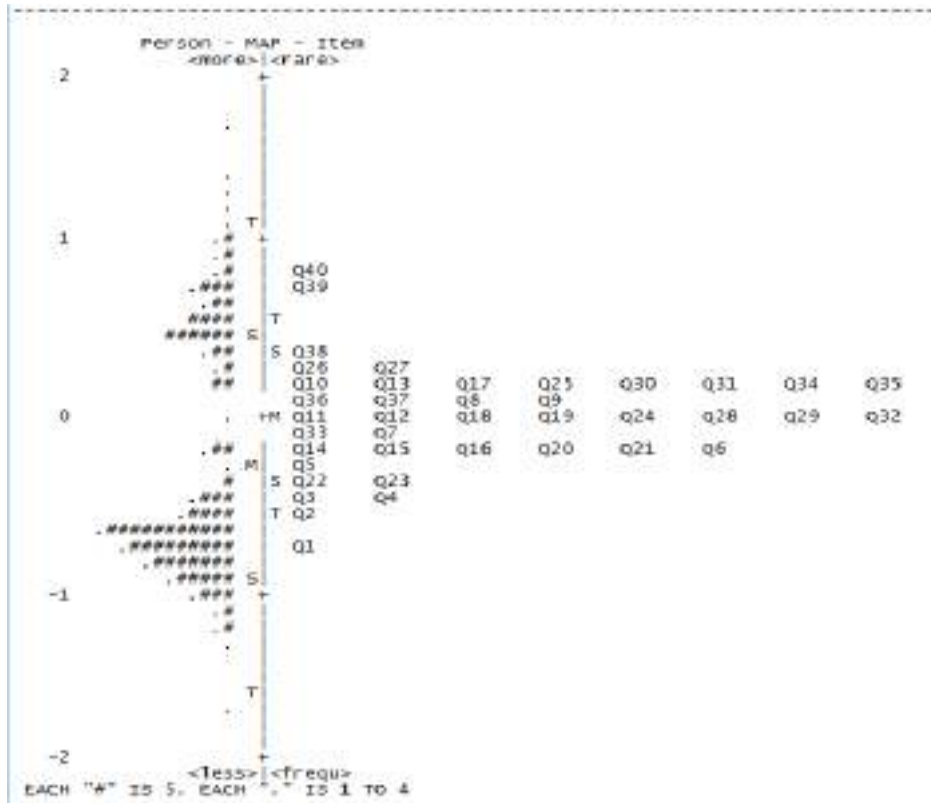


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (column PT-Measure Corr). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than using classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

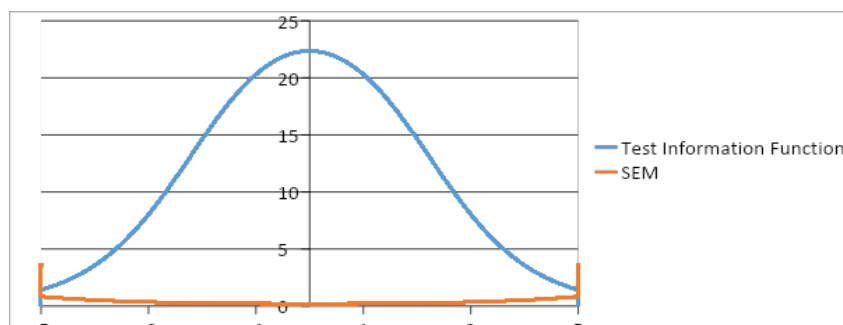


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in vocational school based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

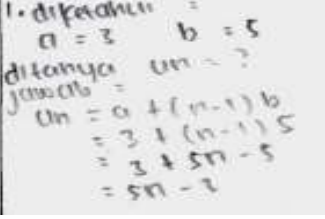
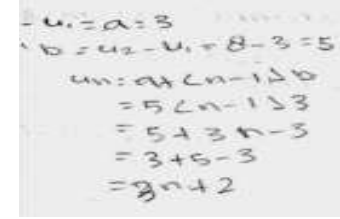
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, ... The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	 <p>1. diketahui $a = 3$ $b = 5$ ditanya $U_n = ?$ jawab $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	 <p>- $u_1 = a = 3$ $b = u_2 - u_1 = 8 - 3 = 5$ $u_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

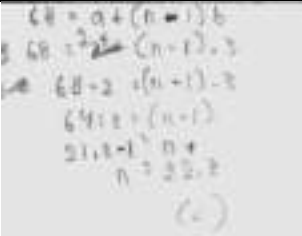
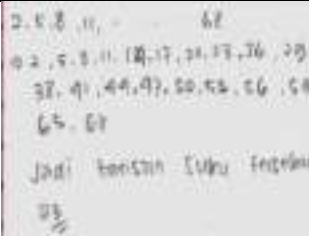
Question2:	Pattern 1	Pattern 2
<p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>		

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

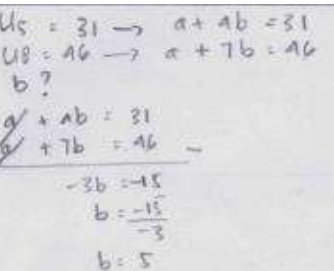
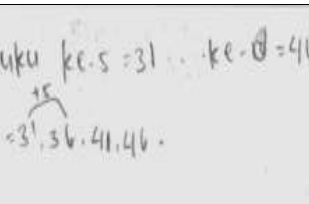
Question 3:	Pattern 1	Pattern 2
<p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>		

Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the n th term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the n th term.

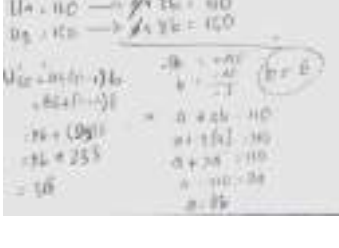
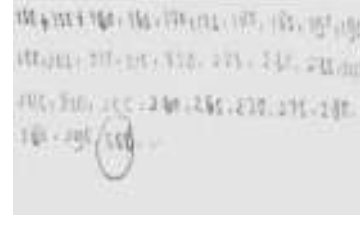
Question 4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>		

Figure 13. An Example of Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

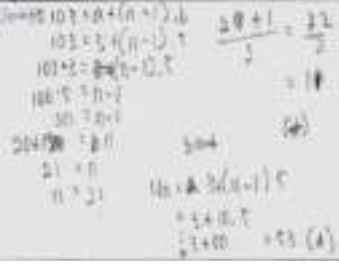

Question 5:	Answer Pattern 1	Answer Pattern 2
<p>An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...308</p> <p>A. 53 B. 52 C. 20 D. 11 E. 10</p> <p>Reason:</p>		

Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice test are easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score, then the result is multiplied by 10 to get a value range of 0–10, or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8, 56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the

teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomus response test is more secondhand and complete than other assessment instruments, such as multiple choice test(Gierl et al., 2017) or essay test(Putri et al., 2020).

Discussion

This research is a development research aimed at developing a good mathematical assessment instrument using polytomic responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to modern theory before revising or replacing the item.

Research on learning assessment with open response politomus was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test, namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

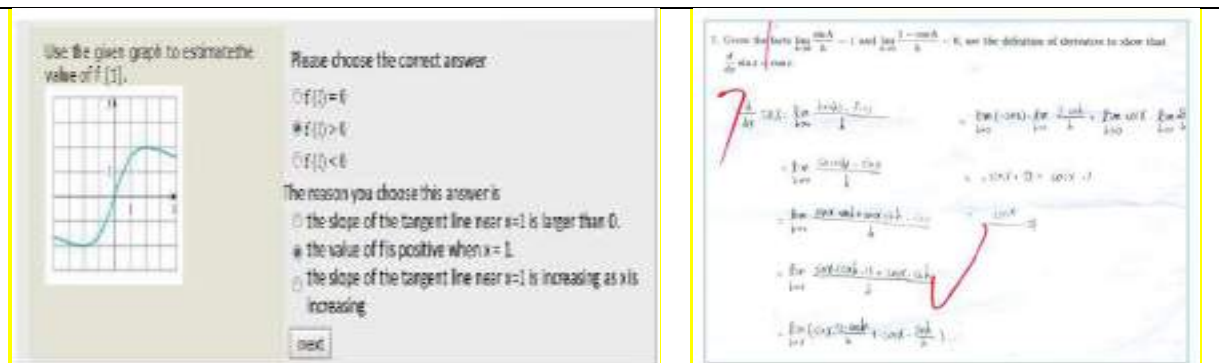


Figure 15. An Example of Student Answers on The Two-Tier Test and

Another study on learning assessment with open response polytomus was conducted by Ayanwale (2021). Ayanwale's research compares two polytomous response test analysis methods, namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

There are other studies related to classical and modern theory, namely Sarea (2018) and Saepuzaman et al. (2021). Sarea's research states that the response polytomus test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomus test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas, and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and

students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

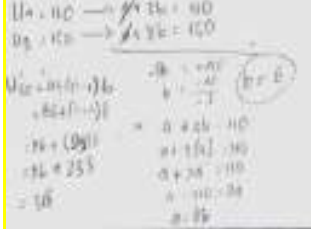
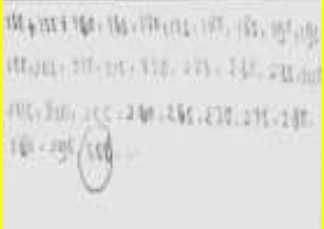
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>	 <p>(i)</p>	 <p>(ii)</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, use a drawing or model, analogy, and formula. Syahlan (2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syahlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching a material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The

description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, conclusions are obtained: (1) the open polytomous response test has a good category according to classical and modern theory. So, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2),95-113. <https://journal.unhas.ac.id/index.php/jmsk/article/view/3402/1939>
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *JurnalSuluh Pendidikan*,17(1),32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan*[Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69.<https://ejournal.upsi.edu.my/index.php/EJSMT/article/view/5029>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: the cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>

- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. [https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_\(1\)%20Lorna.pdf](https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_(1)%20Lorna.pdf)
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geofrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. <https://www.researchgate.net/publication/351451835>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082–1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-

- tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1177/014662168200600401>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on mathematics learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Equating the combined dichotomous and polytomous item test model in an achievement test. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>.
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how?. In A. W. Subianto (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. http://seminar.uny.ac.id/icriems/sites/seminar.uny.ac.id/icriems/files/proceeding2018/M_E69_Asma%20Khiyarunnisa.pdf
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://usnsj.com/index.php/JME/article/view/1607>
- Linacre, J. M. (2012). *Winstep : Rasch-model computer programs*. Winsteps.Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Eirlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. University of Sebelas Maret. <https://jurnal.uns.ac.id/ijscs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Content standards for

primary and secondary education]. Indonesian Government publication service. [https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud Tahun2016 Nomor021 Lampiran.pdf](https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud_Tahun2016_Nomor021_Lampiran.pdf)

- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://journal.unnes.ac.id/sju/index.php/jere/article/view/46133>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://journal.unnes.ac.id/sju/index.php/ujme/article/view/12643>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163. <http://dx.doi.org/10.30587/postulat.v1i2.2094>
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst : Jurnal Pendidikan Fisika dan Terapannya/Karst: Journal of Physics Education and Its Application*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article%20/view/40>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: teori tes klasik and respon [Characteristics of items: classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>

- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*[Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <http://dx.doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran*[Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep*[Analysis of research data with rash and winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122-1?inline=1>
- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <http://jurnal.untad.ac.id/jurnal/index.php/JEPMT/article/view/14181>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian*[Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :
 Class/Department :
 School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
 (use another piece of paper to write down your reason)

- | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>1. Given an arithmetic sequence: 3, 8, 13, 18,
 The formula for the nth term of the sequence is
 A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
 B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
 C. $U_n = 4n - 1$</p> <p>Reason:</p> | <p>2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.
 The number of terms in the sequence is...
 A. 12 D. 23
 B. 13 E. 24
 C. 22</p> <p>Reason:</p> |
| <p>3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
 A. 5 D. 8
 B. 6 E. 11
 C. 7</p> <p>Reason:</p> | <p>4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308 D. 344
 B. 318 E. 354
 C. 326</p> <p>Reason:</p> |
| <p>5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
 Then the middle term of the sequence is ...
 A. 53 D. 11
 B. 52 E. 10
 C. 20
 D. 11
 E. 10</p> <p>Reason:</p> | <p>6. Given the arithmetic sequence: 4, 10, 16, 22,
 If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
 A. 18 D. 24
 B. 20 E. 26
 C. 22</p> <p>Reason:</p> |
| <p>7. The nth term of an arithmetic series is $U_n = 3n - 5$.
 The formula for the sum of the first n terms of the series is ...
 A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
 B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
 C. $S_n = \frac{n}{2}(3n - 4)$</p> <p>Reason:</p> | <p>8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....
 A. $5n - 20$ D. $2n - 20$
 B. $5n - 10$ E. $2n - 10$
 C. $2n - 30$</p> <p>Reason:</p> |
| <p>9. The sum of all integers between 100 and 300 which are divisible by 5 is ...$S_n = \frac{n}{2}(3n - 7)$
 A. 8.200 D. 7.600
 B. 8.000 E. 7.400
 C. 7.800</p> <p>Reason:</p> | <p>10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
 A. 24 D. 27
 B. 25 E. 28
 C. 26</p> <p>Reason:</p> |
| <p>11. The middle term of an arithmetic sequence is 25. If the difference is 4 and the 5th term is 21.
 Then the sum of all the terms in the sequence is ...</p> | <p>12. A number of candies are distributed among five children according to the rules of an arithmetic sequence. The younger the child, the more candy he</p> |

- A. 175
B. 189
C. 275
- D. 295
E. 375

Reason:

gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces

- A. 60
B. 65
C. 70
- D. 75
E. 80

Reason:

13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...
- A. 564
B. 276
C. 48
- D. 45
E. 36

Reason:

14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...
- A. 9
B. 10
C. 11
- D. 12
E. 13

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...
- A. 32
B. 64
C. 128
- D. 256
E. 512

Reason:

16. The value of the middle term of the geometric sequence: 6, 3, ..., $3/512$ is ...

- A. $\frac{1}{16}$
B. $\frac{2}{16}$
C. $\frac{3}{16}$
- D. $\frac{4}{16}$
E. $\frac{5}{16}$

Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm
- A. 18
B. 24
C. 27,5
- D. 35
E. 40,5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...
- A. $\frac{3}{4}$
B. $\frac{1}{4}$
C. $\frac{1}{3}$
- D. $-\frac{1}{2}$
E. $-\frac{3}{4}$

Reason:

19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is.... m
- A. 60
B. 70
C. 80
- D. 90
E. 100

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.

The value of $3a + b$ is ...

- A. 8
B. 10
C. 12
- D. 14
E. 20

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$.

If $K = L$, then c is ...

- A. 12
B. 13
C. 14
- D. 15
E. 16

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$.

Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$
B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$
C. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$
- D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah ...
- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 1 & 3 & 1 \\ 2 & 4 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 5 & 6 \\ 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah ...
- A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...
- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ...
- A. 0 D. 2
 B. 1 E. 4
 C. 2

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$. Inverse matrix AB adalah $(AB)^{-1} = \dots$
- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$

Reason:

33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...
- A. -2 dan $\frac{5}{6}$ D. -2 dan $-\frac{6}{5}$
 B. 2 dan $-\frac{5}{6}$ E. -2 dan $\frac{6}{5}$
 C. 2 dan $\frac{6}{5}$

Reason:

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.
 If $A + B = C$, then $x + y = \dots$
- A. -5 D. 3
 B. -1 E. 5
 C. 1

Reason:

26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$ then $A(B - C) = \dots$
- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix. Value of x that satisfies is ...
- A. -5 D. 3
 B. -4 E. 4
 C. -3

Reason:

30. Transpose matrix P adalah P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ...
- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$

Reason:

32. The roots of the quadratic equation $x^2 - 3x + 4 = 0$ are ...
- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$

Reason:

34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...
- A. -4 D. 2
 B. -2 E. 4
 C. 0

Reason:

Sugeng Sutiarmo

Lampung University

[Quoted text hidden]

2 attachments



Cert of proofreading_Sugeng Sutiarmo.pdf

259K



5th Revision_Article Sugeng Sutiarmo et al.docx

1689K



KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS LAMPUNG
UPT BAHASA

Jalan Prof. Dr. Soemantri Brojonegoro No. 1 Bandar Lampung 35145
Telepon : (0721) 770844, Whatsapp : 0811 724 5544, email : uptbahasa@kpa.unila.ac.id
Website : uptbahasa.unila.ac.id



CERTIFICATE OF PROOFREADING

Number: **105** /UN26.33/TU.00.08/2022

The undersigned below,

Name : Dr. Muhammad Sukirlan, M.A.

NIP : 196412121990031003


Position : Head of Language Center - University of Lampung

states that the article entitled : **"Developing Assessment Instrument Using Polytomous Response in Mathematics"** written by **Sugeng Sutiarmo, Undang Rosidin, Aan Sulistiawan** has been edited or proofread in terms of linguistic aspects by Language Center University of Lampung.



Bandar Lampung, April 19th, 2022

Head,


Dr. Muhammad Sukirlan, M.A.
NIP 196412121990031003

Developing Assessment Instrument Using Polytomous Response in Mathematics

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) open polytomous response test has a good category according to classical and modern theory. However the discrimination power of test items in classical theory need several revisions, (2) the assessment instrument using polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. Conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm,

assessment in schools is divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).

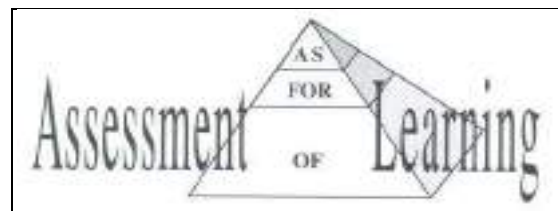


Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with certain rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over essay test are firstly, the can be conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weakness. Multiple-choice test is not able to potray the actual abilities of students, the answers of this test tend to be guessing game or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). Multiple-choice test with only two answer choices are called "dichotomous test," and multiple-choice test with more than two answer choices are called "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the

weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or hence forth called as polytomous response test (Suwanto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as, test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into open polytomous response test. Open polytomous response test is a form of multiple-

choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response test for students in college, and senior or junior school. Students in college, and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools is not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomus test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmawati, 2020). Therefore the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomus response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations,

and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows: (1) Does the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory?, and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's model (Plomp, 2013), with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

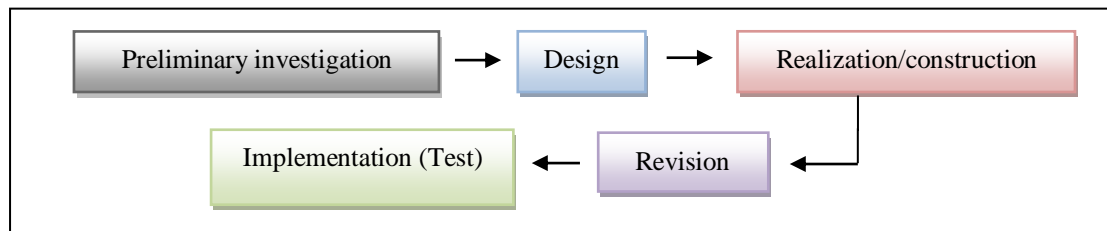


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose

mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. *Questionnaire data analysis (qualitative analysis)*

There are two sets of questionnaire data namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. *Test data analysis (empirical analysis)*

After conducting the content validity test, the researchers conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to having good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the IteMan program was used for classical theory and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 *Analysis of test data with classical theory*

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.

- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Corr column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional

solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Corr is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay test and 20% of teachers used multiple-choice test, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as: teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

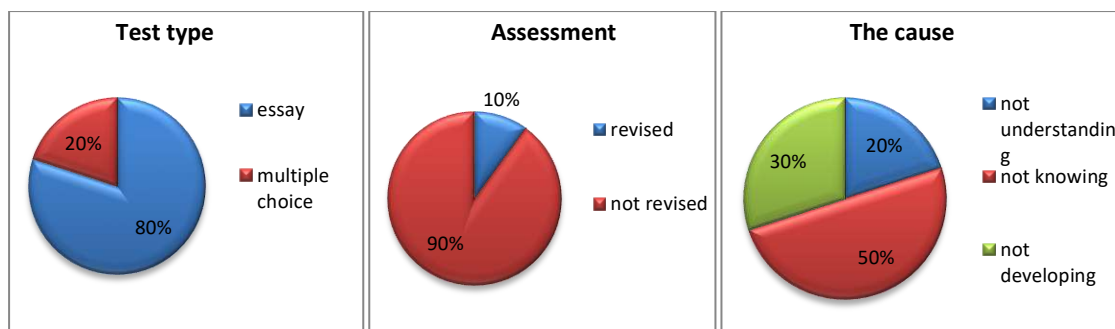


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's Alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised

4	0.540	good	0.304	Good	24	0.438	good	-0.071	Revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	Revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	Revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	Revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	Revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	Revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	Revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	Revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	Revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	Revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	Revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	Revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	Revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	Revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	Revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	Revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

Analysis of Test Data with Modern Theory

The Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first factor Eigenvalue is greater than

20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

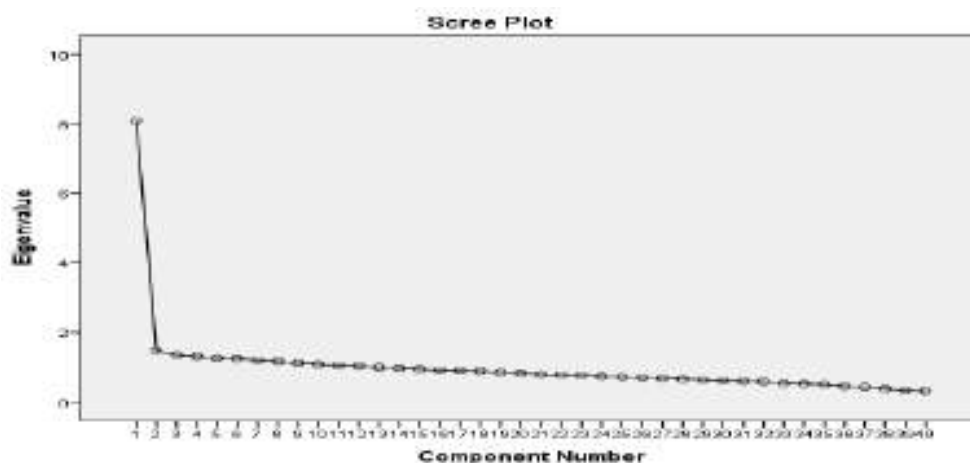


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be

determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	<i>K1</i>	<i>K2</i>	<i>K3</i>	<i>K4</i>	<i>K5</i>	<i>K6</i>	<i>K7</i>	<i>K8</i>	<i>K9</i>	<i>K10</i>
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model". If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Corr is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MOOGL S. E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT MATCH OBS%	EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.43	.07	.97	-1.4	.98	-1.5	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-3.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.8	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.3	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.9	.97	-.5	.45	.44	48.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	964	413	-.01	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.3	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.3	1.08	1.2	.56	.44	43.6	50.6	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	-.16	.07	.97	-.4	.97	-.4	.31	.44	50.0	50.9	Q26
27	936	413	-.20	.07	1.02	-.3	1.02	-.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	-.14	.07	1.06	.9	1.01	.9	.30	.44	45.5	50.9	Q30
31	958	413	-.11	.07	1.07	1.1	1.07	1.1	.35	.44	48.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.04	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	-.09	.07	1.06	1.0	1.07	1.1	.35	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.69	-6.3	.66	-6.2	.39	.43	38.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained was be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

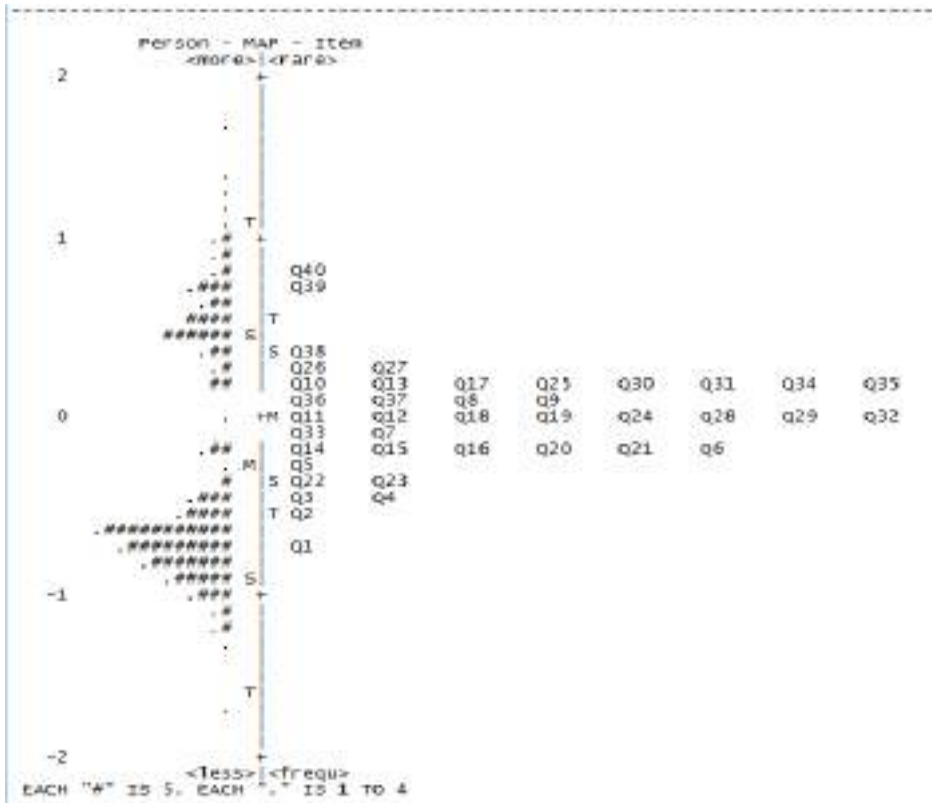


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (column PT-Measure Corr). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than using classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

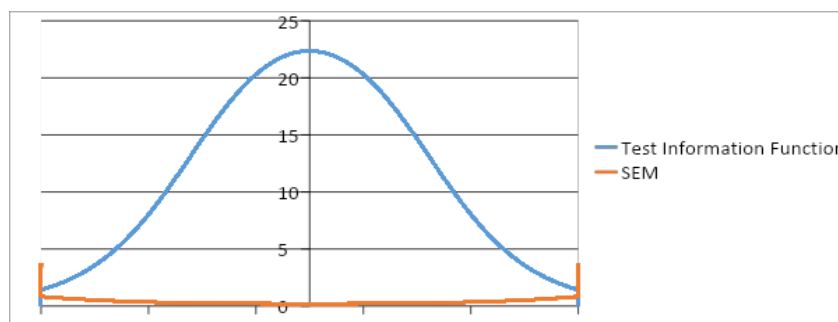


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in vocational school based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

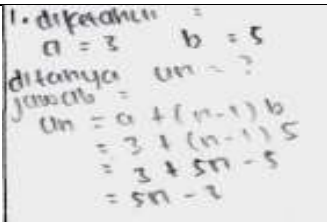
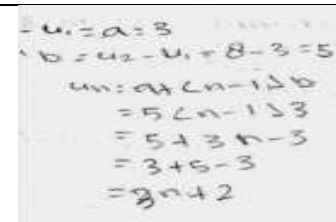
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, ... The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	 <p>1. diketahui $a = 3$ $b = 5$ ditanya $U_n = ?$ jawab $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	 <p>$U_n = a + (n-1)b$ $U_n = 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

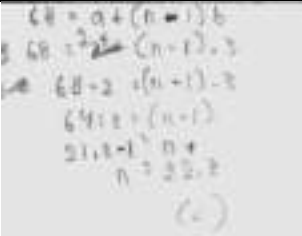
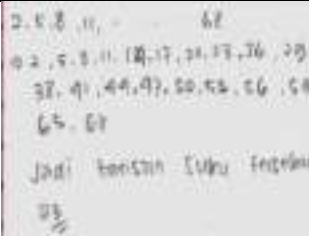
Question2:	Pattern 1	Pattern 2
Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is... A. 12 B. 13 C. 22 D. 23 E. 24 Reason:	 $U_n = a + (n-1)b$ $68 = 2 + (n-1) \cdot 3$ $68 - 2 = (n-1) \cdot 3$ $66 = 3(n-1)$ $22 = n-1$ $n = 22 + 1$ $n = 23$	 <p>2, 5, 8, 11, ..., 68</p> <p>02, 05, 08, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 59, 62, 65, 68</p> <p>Jadi banyak suku adalah 23</p>

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

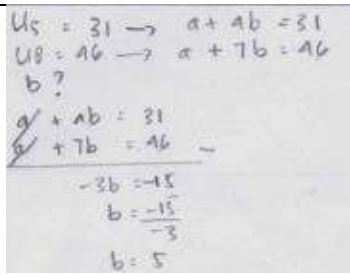
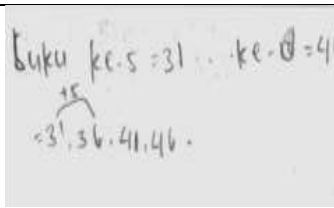
Question 3:	Pattern 1	Pattern 2
An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is... A. 5 B. 6 C. 7 D. 8 E. 11 Reason:	 $U_5 = 31 \rightarrow a + 4b = 31$ $U_8 = 46 \rightarrow a + 7b = 46$ $b = ?$ $\begin{array}{r} a + 4b = 31 \\ a + 7b = 46 \\ \hline -3b = -15 \\ b = \frac{-15}{-3} \\ b = 5 \end{array}$	 <p>buku ke-5 = 31 .. ke-8 = 46</p> <p>31, 36, 41, 46</p>

Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the n th term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the n th term.

Question 4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>		

Figure 13. An Example of Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

Question 5:	Answer Pattern 1	Answer Pattern 2
<p>An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...308</p> <p>A. 53 B. 52 C. 20 D. 11 E. 10</p> <p>Reason:</p>		

Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test

and an essay. Multiple choice test are easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score, then the result is multiplied by 10 to get a value range of 0–10, or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through

the open polytomus response test is more secondhand and complete than other assessment instruments, such as multiple choice test (Gierl et al., 2017) or essay test (Putri et al., 2020).

Discussion

This research is a development research aimed at developing a good mathematical assessment instrument using polytomic responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to modern theory before revising or replacing the item.

Research on learning assessment with open response politomus was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test, namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

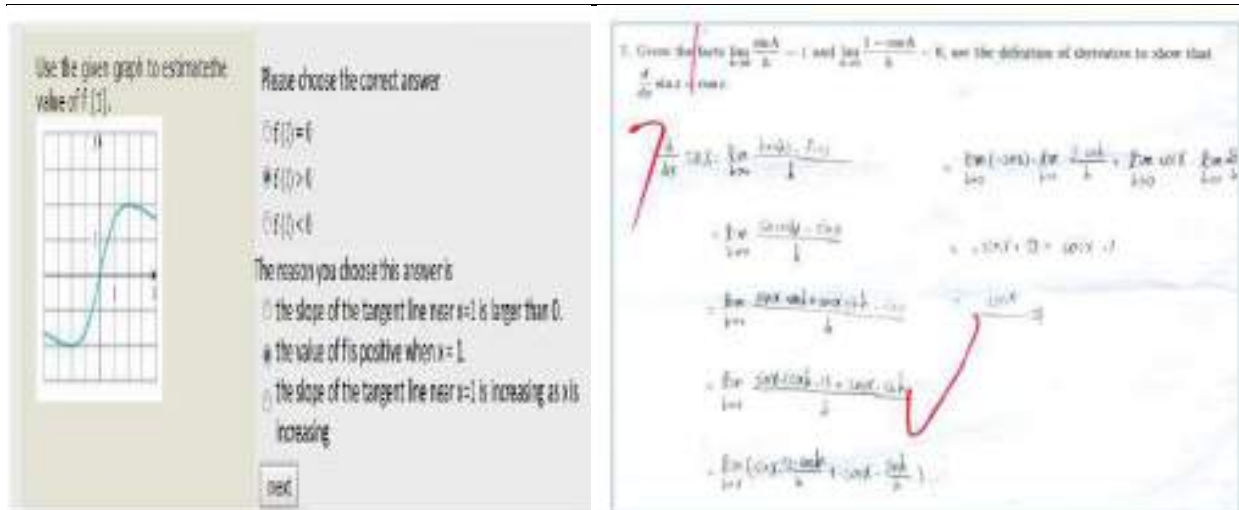


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous Response Test

Another study on learning assessment with open response polytomus was conducted by Ayanwale (2021). Ayanwale's research compares two polytomous response test analysis methods, namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

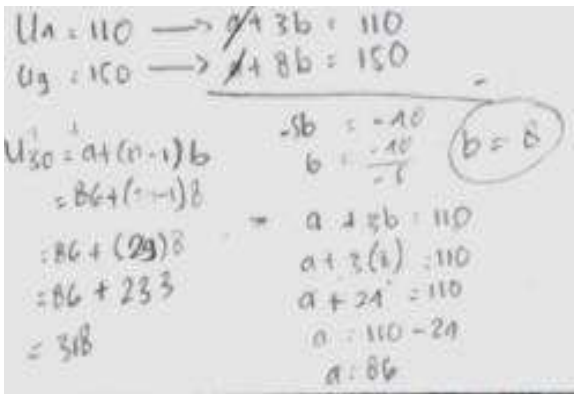
Other studies related to classical and modern theory conducted by Sarea (2018) and Saepuzaman et al.(2021). Sarea's research states that the response polytomus test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomus test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas, and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

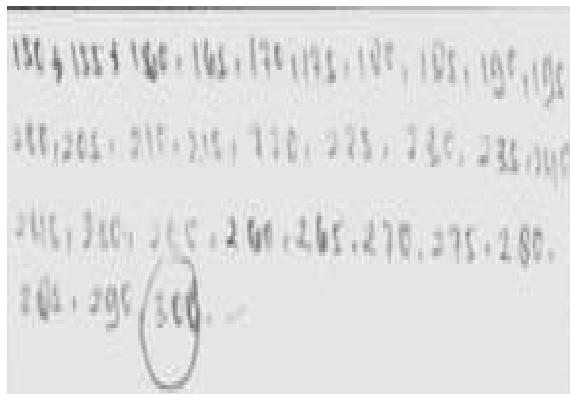
The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

A. 308
B. 318
C. 326
D. 344
E. 354

Reason:



(i)



(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, use a drawing or model, analogy, and formula. Syahlan (2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syahlan's opinion,

students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching a material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. So, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://journal.unhas.ac.id/index.php/jmsk/article/view/3402/1939>

- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *Jurnal Suluh Pendidikan*, 17(1), 32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://ejournal.upsi.edu.my/index.php/EJSMT/article/view/5029>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: the cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. [https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_\(1\)%20Lorna.pdf](https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_(1)%20Lorna.pdf)
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geofrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. <https://www.researchgate.net/publication/351451835>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of*

- Educational Research*, 87(6), 1082–1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar [Equalization of the dichotomous and polytomus mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how?. In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. http://seminar.uny.ac.id/icriems/sites/seminar.uny.ac.id/icriems/files/proceeding2018/ME69_Asma%20Khiyarunnisa.pdf
- Khusnah, M.(2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148.

<https://usnsj.com/index.php/JME/article/view/1607>

- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://jurnal.uns.ac.id/ijscs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Standards of content for primary and secondary education]. Indonesian Government publication service. [https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud Tahun2016 Nomor021.pdf](https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud_Tahun2016_Nomor021.pdf)
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://journal.unnes.ac.id/sju/index.php/jere/article/view/46133>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://journal.unnes.ac.id/sju/index.php/ujme/article/view/12643>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163.

<http://dx.doi.org/10.30587/postulat.v1i2.2094>

- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst : Jurnal Pendidikan Fisika dan Terapannya/Karst: Journal of Physics Education and Its Application*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article%20view/40>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: teori tes klasik and respon [Characteristics of items: classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <http://dx.doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep* [Analysis of research data with rash and winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122->

[1?inline=1](#)

- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <http://jurnal.untad.ac.id/jurnal/index.php/JEPMT/article/view/14181>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :
 Class/Department :
 School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
 (use another piece of paper to write down your reason)

- | | |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>1. Given an arithmetic sequence: 3, 8, 13, 18,
 The formula for the nth term of the sequence is
 A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
 B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
 C. $U_n = 4n - 1$</p> <p>Reason:</p> | <p>2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.
 The number of terms in the sequence is...
 A. 12 D. 23
 B. 13 E. 24
 C. 22</p> <p>Reason:</p> |
| <p>3. An arithmetic sequence, the 5th term is 31 and the
 8th term is 46. The difference between the
 sequences is...
 A. 5 D. 8
 B. 6 E. 11
 C. 7</p> <p>Reason:</p> | <p>4. The 4th and 9th terms of an arithmetic sequence are
 110 and 150. The 30th terms of the arithmetic
 sequence are...
 A. 308 D. 344
 B. 318 E. 354
 C. 326</p> <p>Reason:</p> |
| <p>5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
 Then the middle term of the sequence is ...
 A. 53 D. 11
 B. 52 E. 10
 C. 20
 D. 11
 E. 10</p> <p>Reason:</p> | <p>6. Given the arithmetic sequence: 4, 10, 16, 22,
 If two numbers are inserted in every two
 consecutive terms then the 10th term of the sequence
 is...
 A. 18 D. 24
 B. 20 E. 26
 C. 22</p> <p>Reason:</p> |
| <p>7. The nth term of an arithmetic series is $U_n = 3n - 5$.
 The formula for the sum of the first n terms of the
 series is ...
 A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n =$
 $\frac{n}{2}(3n - 3)$
 B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
 C. $S_n = \frac{n}{2}(3n - 4)$</p> <p>Reason:</p> | <p>8. The sum of the first n terms of an arithmetic series.
 $S_n = n^2 - 19n$. The formula for the nth term of the
 Series is.....
 A. $5n - 20$ D. $2n - 20$
 B. $5n - 10$ E. $2n - 10$
 C. $2n - 30$</p> <p>Reason:</p> |
| <p>9. The sum of all integers between 100 and 300 which
 are divisible by 5 is ...$S_n = \frac{n}{2}(3n - 7)$
 A. 8.200 D. 7.600
 B. 8.000 E. 7.400
 C. 7.800</p> <p>Reason:</p> | <p>10. PT. Angkasa Jaya in the first year produced 5,000
 units. In the second year, production was reduced
 by 80 units per year. In what year did the company
 produce 3,000 units?
 A. 24 D. 27
 B. 25 E. 28
 C. 26</p> <p>Reason:</p> |

11. The middle term of an arithmetic sequence is 25. If the difference is 4 and the 5th term is 21.

Then the sum of all the terms in the sequence is ...

- A. 175
- B. 189
- C. 275
- D. 295
- E. 375

Reason:

12. A number of candies are distributed among five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces

- A. 60
- B. 65
- C. 70
- D. 75
- E. 80

Reason:

13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...

- A. 564
- B. 276
- C. 48
- D. 45
- E. 36

Reason:

14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...

- A. 9
- B. 10
- C. 11
- D. 12
- E. 13

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...

- A. 32
- B. 64
- C. 128
- D. 256
- E. 512

Reason:

16. The value of the middle term of the geometric sequence: 6, 3, ..., $3/512$ is ...

- A. $\frac{1}{16}$
- B. $\frac{1}{16}$
- C. $\frac{3}{16}$
- D. $\frac{4}{16}$
- E. $\frac{1}{5}$

Reason:.....

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm

- A. 18
- B. 24
- C. 27,5
- D. 35
- E. 40,5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...

- A. $\frac{3}{4}$
- B. $\frac{1}{4}$
- C. $\frac{1}{3}$
- D. $-\frac{1}{2}$
- E. $-\frac{3}{4}$

Reason:

19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is... m

- A. 60
- B. 70
- C. 80
- D. 90
- E. 100

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.

The value of $3a + b$ is ...

- A. 8
- B. 10
- C. 12
- D. 14
- E. 20

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L =$

$$\begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$$

If $K = L$, then c is ...

- A. 12
- B. 13
- C. 14
- D. 15
- E. 16

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$.

Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$
- C. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$
- D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah ...
- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 1 & 3 & 1 \\ 0 & 7 & 6 \\ 2 & 4 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 5 & 6 \\ 1 & 3 & 1 \\ 2 & 4 & 5 \end{bmatrix}$
24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.
 If $A + B = C$, then $x + y = \dots$
 A. -5 D. 3
 B. -1 E. 5
 C. 1

Reason:

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah ...
- A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$
26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$
 then $A(B - C) = \dots$

Reason:

- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...
- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$
28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix.
 Value of x that satisfies is ...

Reason:

- A. -5 D. 3
 B. -4 E. 4
 C. -3

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ...
- A. 0 D. 2
 B. 1 E. 4
 C. 2
30. Transpose matrix P adalah P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ...

Reason:

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$.
 Inverse matrix AB adalah $(AB)^{-1} = \dots$
- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$
32. The roots of the quadratic equation $3x^2 - 4x - 4 = 0$ are ...

Reason:

Reason:

33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...
- A. -2 dan $\frac{5}{6}$ D. -2 dan $-\frac{6}{5}$
 B. 2 dan $-\frac{5}{6}$ E. -2 dan $\frac{6}{5}$
 C. 2 dan $\frac{6}{5}$
34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...

Reason:

Reason:

35. The roots of the quadratic equation $2x^2 - 3x - 9 = 0$ are x_1 dan x_2 . Value of $x_1^2 + x_2^2$ is ...

- A. $11\frac{1}{4}$
- B. $6\frac{3}{4}$
- C. $2\frac{1}{4}$
- D. $-6\frac{3}{4}$
- E. $-11\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...

- A. $y = x^2 - 2x + 1$
- B. $y = x^2 - 2x + 3$
- C. $y = x^2 + 2x - 1$
- D. $y = x^2 + 2x + 1$
- E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0, -4)$ then the value of $f(7)$ is ...

- A. -16
- B. -17
- C. -18
- D. -19
- E. -20

Reason:

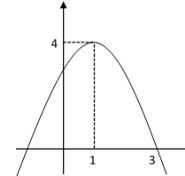
36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α dan β . The quadratic equation $(\alpha - 2)$ dan $(\beta - 2)$ is ...

- a. $x^2 + 6x + 5 = 0$
- b. $x^2 + 6x + 7 = 0$
- c. $x^2 + 6x + 11 = 0$
- d. $x^2 - 2x + 3 = 0$
- e. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation ? ...

- A. $y = x^2 + 2x + 3$
- B. $y = x^2 - 2x - 3$
- C. $y = -x^2 + 2x - 3$
- D. $y = -x^2 - 2x + 3$
- E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...

- A. $y = -x^2 + 2x - 3$
- B. $y = -x^2 + 2x + 3$
- C. $y = -x^2 - 2x + 3$
- D. $y = -x^2 - 2x - 5$
- E. $y = -x^2 - 2x + 5$

Reason:



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Acceptance Letter for the Manuscript ID# 21112502244011

3 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>

Wed, Apr 20, 2022 at 6:21 PM

To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Dear Dr. Sugeng Sutiarso,

Congratulation! After a thorough double-blind review, I am pleased to inform you that your manuscript entitled "*Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School*" (ID#21112502244011) has been accepted. It is scheduled for publication in the Volume 11 Issue 3 of the "*European Journal of Educational Research*".

We kindly ask you to pay the article processing fee USD 600 [+USD 50 transaction fee of the receiver bank] **totally USD 650** via bank wire transfer. Kindly acknowledge invoice of this acceptance letter. Payment due date: **April 22, 2022**.

BANK WIRE TRANSFER INFORMATION :

NAME OF BENEFICIARY:	Ahmet Cezmi SAVAŞ
ADDRESS OF BENEFICIARY:	Degirmicem District Ozgurluk Str. No:32B , Zipcode:27090, Gaziantep, TURKEY
PHONE OF BENEFICIARY:	+90 (342) 909 61 90
CORRESPONDENT BANK CHARGER:	REMITTER
AMOUNT:	USD 650
PAYMENT DETAIL:	EU-JER_ Manuscript ID# 21112502244011
BANK NAME:	QNB Finansbank
BANK ADDRESS:	Esentepe Mahallesi Büyükdere Caddesi Kristal Kule Binası No:215 Şişli - İstanbul
BRANCH OF THE BANK:	ENPARA
BRANCH CODE:	3663
ACCOUNT NUMBER:	88177946
IBAN:	TR66 0011 1000 0000 0088 1779 46
SWIFT CODE:	FNNBTRISXXX

After payment, we will send the gallery proof of your paper. The galley proofs must be returned to us within 2 calendar days. Furthermore, you are responsible for any error in the published paper due to your oversight.

Please let us know, when you get this email. We looking forward to getting your payment in order to continue the editorial process.

PS: Please do the attached additional minor corrections and send your finalized paper in 2 days.

Best regards.

Ahmet C. Savas Ph.D.

Editor, European Journal of Educational Research

<http://www.eu-jer.com>

editor@eu-jer.com

2 attachments**MINOR_EU-JER_21112502244011.docx**

1693K

**Acceptance Letter for the EU-JER_ID#21112502244011.pdf**

287K

SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Fri, Apr 22, 2022 at 11:08 AM

Dear. Ahmet C. Savas Ph.D.

Editor, European Journal of Educational Research.

Here I attach a revision of my article according to your suggestion, and proof of remittance for payment of the article processing fee (USD 650).

Note: Information from our bank that the money transfer process or will arrive at you within 3-5 days.

Best regards,
Sugeng Sutiarso
Lampung University

[Quoted text hidden]

2 attachments**Revision_Sugeng Sutiarso_MINOR_EU-JER_21112502244011.docx**

1690K

**Remittance_Sugeng Sutiarso.pdf**

252K

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Fri, Apr 22, 2022 at 3:51 PM

Dear Dr. Sutiarso,

Thank you for your payment. We are waiting for the appearing in our account. It may takes a few days because international money transfer. We will inform you when we get it.

We have received your finalized paper. We have started to prepare the galley proof of your paper.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

[Quoted text hidden]

Developing Assessment Instrument Using Polytomous Response in Mathematics

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) open polytomous response test has a good category according to classical and modern theory. However the discrimination power of test items in classical theory need several revisions, (2) the assessment instrument using polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. Conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm,

assessment in schools is divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).

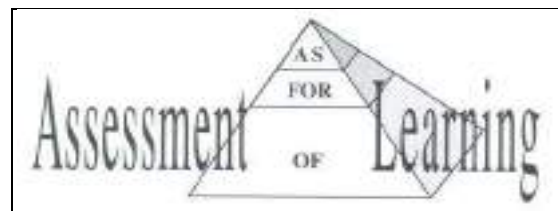


Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with certain rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over essay test are firstly, the test can be conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weaknesses. Multiple-choice test is not able to portray the actual abilities of students, the answers of this test tend to be guessing game or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). Multiple-choice test with only two answer choices is called "dichotomous test," and multiple-choice test with more than two answer choices is called "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the

weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or hence forth called as polytomous response test (Suwanto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as, test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into open polytomous response test. Open polytomous response test is a form of multiple-

choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response test for students in college, and senior or junior school. Students in college, and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools is not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomus test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmawati, 2020). Therefore the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomus response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations,

and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows: (1) Does the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory? and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's model (Plomp, 2013), with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

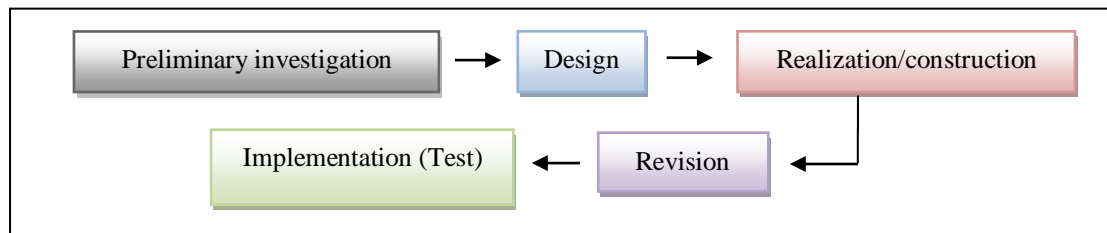


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arifNU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose

mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. *Questionnaire data analysis (qualitative analysis)*

There are two sets of questionnaire data namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. *Test data analysis (empirical analysis)*

After conducting the content validity test, the researchers conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to having good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the IteMan program was used for classical theory and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 *Analysis of test data with classical theory*

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.

- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep), because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Correlation column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional

solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit MNSQ value is 0.5 to 1.5 and the Pt-Measure Correlation is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay test and 20% of teachers used multiple-choice test, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as: teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

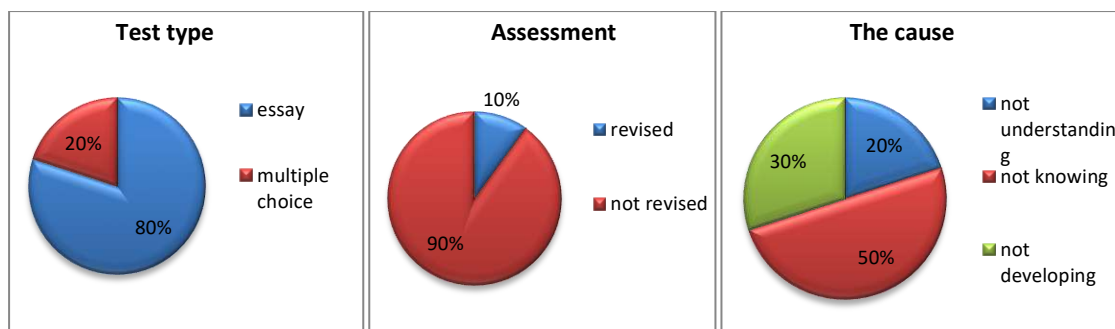


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's Alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised

4	0.540	good	0.304	Good	24	0.438	good	-0.071	Revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	Revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	Revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	Revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	Revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	Revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	Revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	Revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	Revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	Revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	Revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	Revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	Revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	Revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	Revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	Revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

Analysis of Test Data with Modern Theory

The Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first factor Eigenvalue is greater than

20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

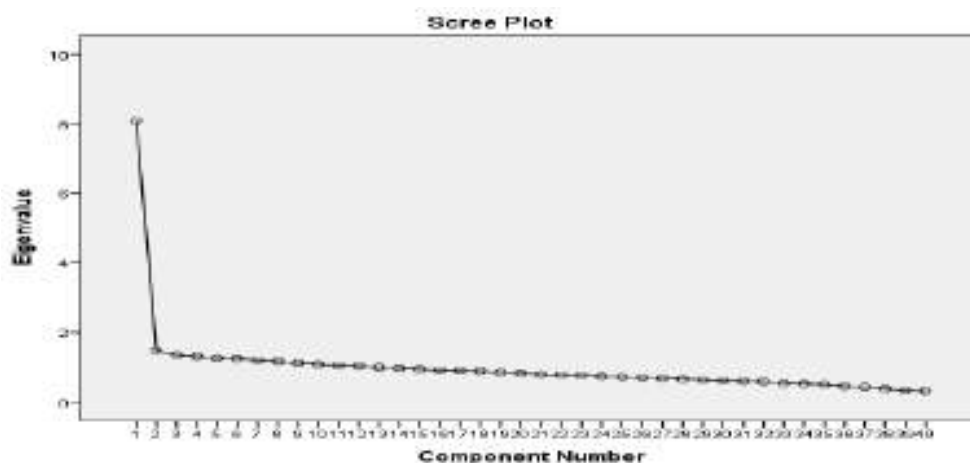


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be

determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	<i>K1</i>	<i>K2</i>	<i>K3</i>	<i>K4</i>	<i>K5</i>	<i>K6</i>	<i>K7</i>	<i>K8</i>	<i>K9</i>	<i>K10</i>
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model". If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Correlation is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MOOGL S. E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT MATCH OBS%	EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.43	.07	.97	-1.4	.98	-1.5	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-3.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.8	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.3	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.9	.97	-.5	.45	.44	48.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	964	413	-.01	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.3	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.3	1.08	1.2	.56	.44	43.6	50.6	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	-.16	.07	.97	-.4	.97	-.4	.31	.44	50.0	50.9	Q26
27	936	413	-.20	.07	1.02	-.3	1.02	-.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	-.14	.07	1.06	.9	1.01	.9	.30	.44	45.5	50.9	Q30
31	958	413	-.11	.07	1.07	1.1	1.07	1.1	.35	.44	48.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.04	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	-.09	.07	1.06	1.0	1.07	1.1	.35	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.69	-6.3	.66	-6.2	.39	.43	38.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained was be presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

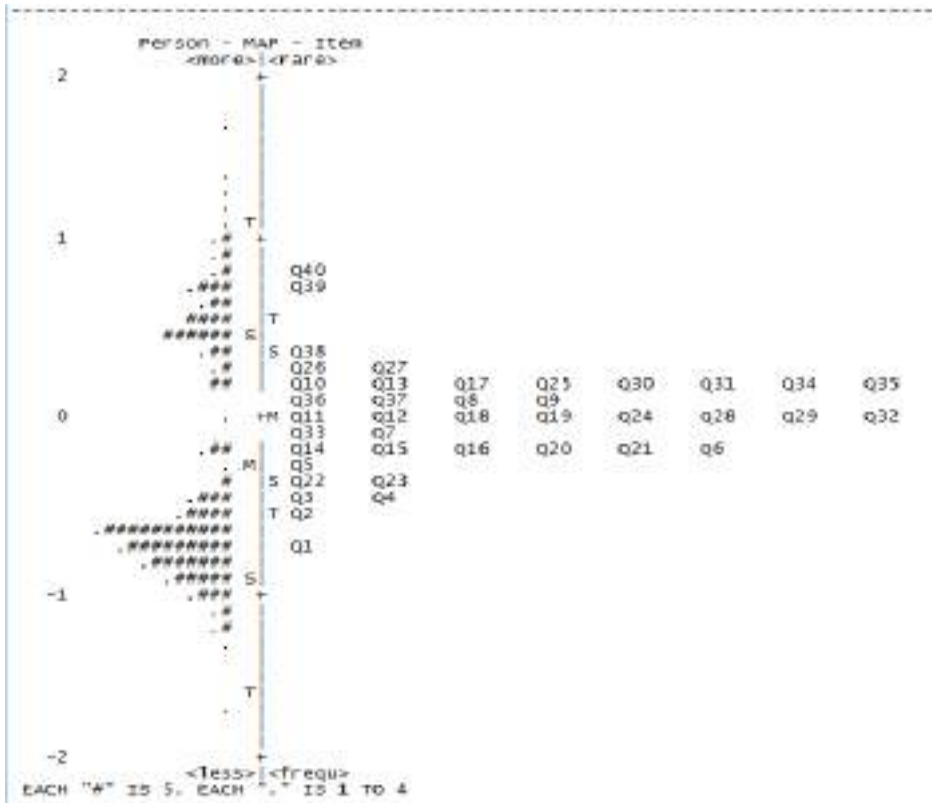


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (PT-Measure Correlation column). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than using classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

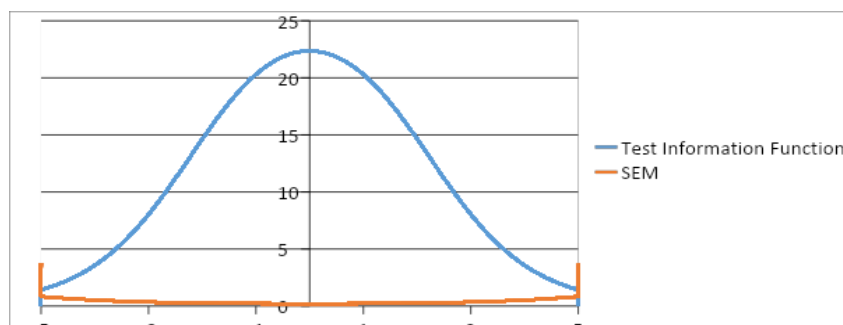


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in vocational school based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

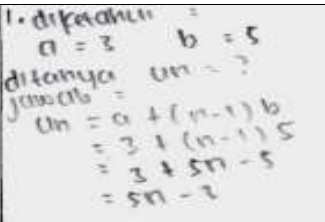
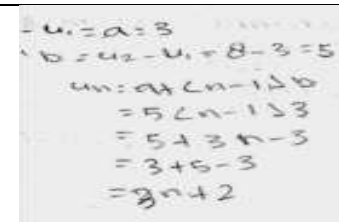
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, ... The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>		

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term), and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

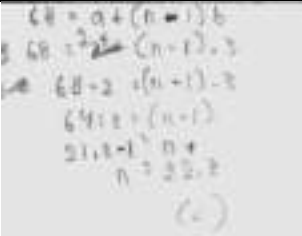
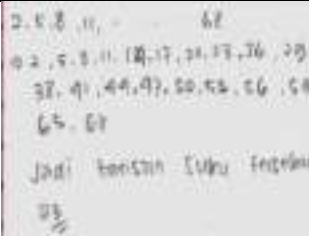
Question2:	Pattern 1	Pattern 2
<p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>	 <p> $68 = a + (n-1)b$ $68 = 2 + (n-1) \cdot 3$ $68 - 2 = (n-1) \cdot 3$ $64 = 3(n-1)$ $21,3 = 1 \cdot n +$ $n = 22,3$ </p>	 <p> 2, 5, 8, 11, ..., 68 02, 05, 08, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 59, 62, 65, 68 Jadi konstan (uku) frekuensi </p>

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

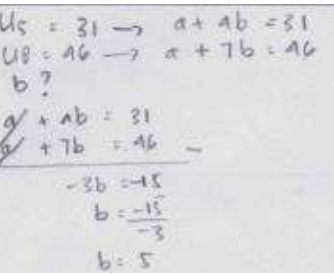
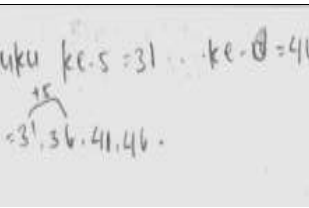
Question 3:	Pattern 1	Pattern 2
<p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>	 <p> $U_5 = 31 \rightarrow a + 4b = 31$ $U_8 = 46 \rightarrow a + 7b = 46$ $b = ?$ $\begin{array}{r} a + 4b = 31 \\ a + 7b = 46 \\ \hline -3b = -15 \\ b = \frac{-15}{-3} \\ b = 5 \end{array}$ </p>	 <p> buku ke-5 = 31 .. ke-8 = 46 $+3$ = 31, 36, 41, 46. </p>

Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the n th term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the n th term.

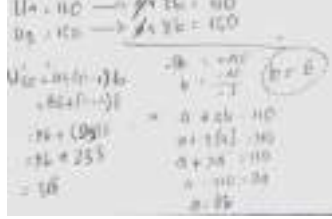
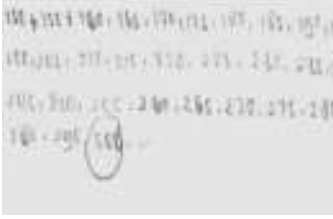
Question 4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>		

Figure 13. An Example of Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

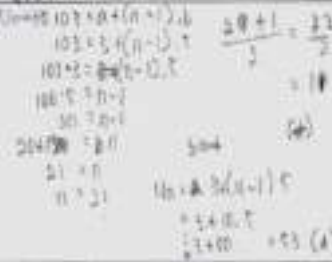
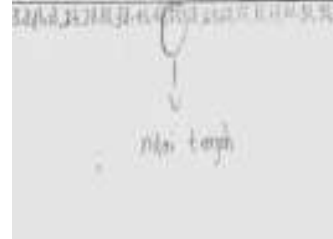
Question 5:	Answer Pattern 1	Answer Pattern 2
<p>An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...308</p> <p>A. 53 B. 52 C. 20 D. 11 E. 10</p> <p>Reason:</p>		

Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test

and an essay. Multiple choice test are easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score, then the result is multiplied by 10 to get a value range of 0–10, or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through

the open polytomus response test is more secondhand and complete than other assessment instruments, such as multiple choice test (Gierl et al., 2017) or essay test (Putri et al., 2020).

Discussion

This research is development research aimed at developing a good mathematical assessment instrument using polytomic responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to modern theory before revising or replacing the item.

Research on learning assessment with open response politomus was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test, namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

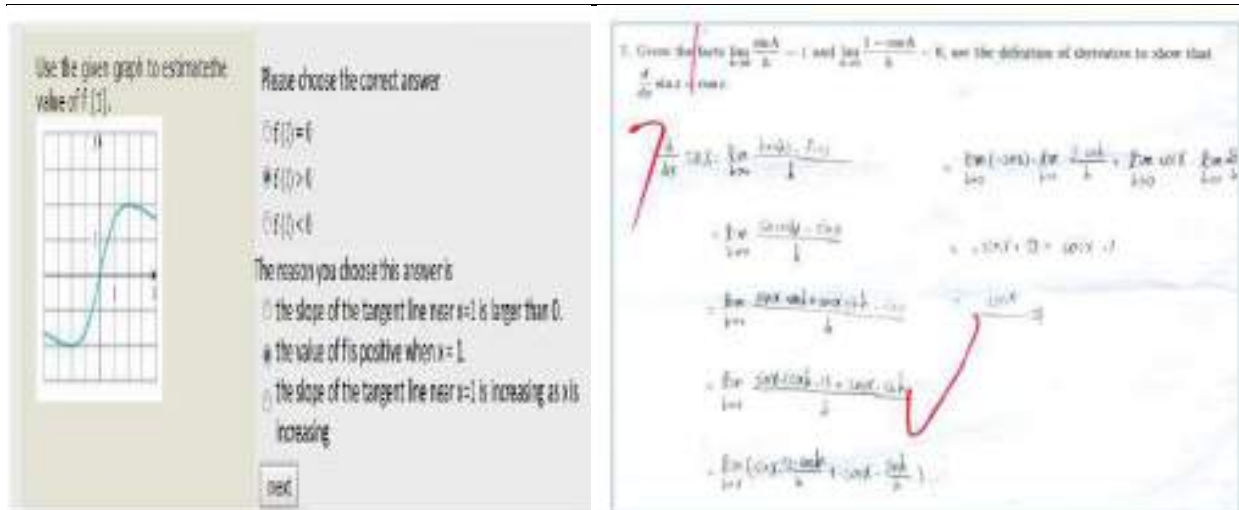


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous Response Test

Another study on learning assessment with open response polytomus was conducted by Ayanwale (2021). Ayanwale's research compares two polytomous response test analysis methods, namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

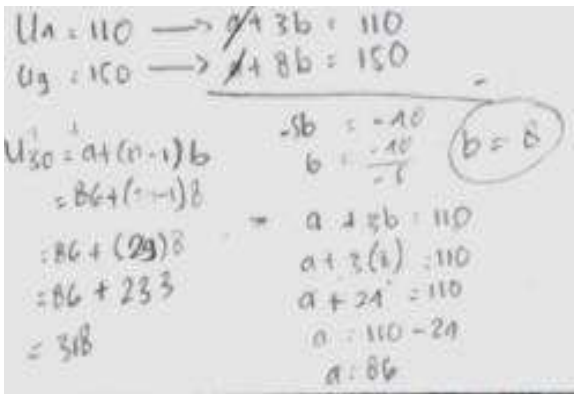
Other studies related to classical and modern theory conducted by Sarea (2018) and Saepuzaman et al.(2021). Sarea's research states that the response polytomus test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomus test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

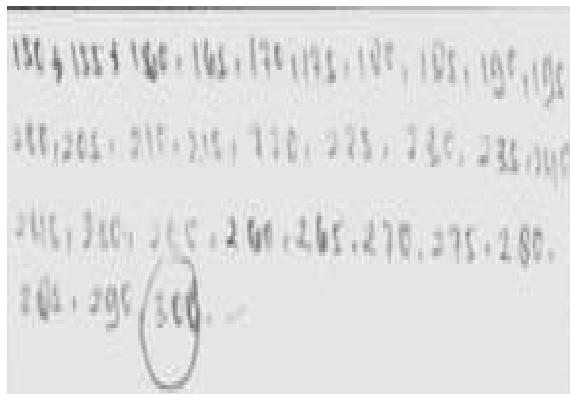
The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

A. 308
B. 318
C. 326
D. 344
E. 354

Reason:



(i)



(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, use a drawing or model, analogy, and formula. Syahlan(2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syahlan's opinion,

students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching a material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. So, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://journal.unhas.ac.id/index.php/jmsk/article/view/3402/1939>

- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *Jurnal Suluh Pendidikan*, 17(1), 32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://ejournal.upsi.edu.my/index.php/EJSMT/article/view/5029>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: the cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement?. The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. [https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_\(1\)%20Lorna.pdf](https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_(1)%20Lorna.pdf)
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geofrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. ResearchGate. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of*

- Educational Research*, 87(6), 1082–1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar [Equalization of the dichotomous and polytomus mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how?. In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. http://seminar.uny.ac.id/icriems/sites/seminar.uny.ac.id/icriems/files/proceeding2018/ME69_Asma%20Khiyarunnisa.pdf
- Khusnah, M.(2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148.

<https://usnsj.com/index.php/JME/article/view/1607>

- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://jurnal.uns.ac.id/ijscs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Standards of content for primary and secondary education]. Indonesian Government publication service. [https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud Tahun2016 Nomor021.pdf](https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud_Tahun2016_Nomor021.pdf)
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://journal.unnes.ac.id/sju/index.php/jere/article/view/46133>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://journal.unnes.ac.id/sju/index.php/ujme/article/view/12643>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha Medika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163.

<http://dx.doi.org/10.30587/postulat.v1i2.2094>

- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst: Journal of Physics Education and Its Application/Karst :Jurnal Pendidikan Fisika dan Terapannya*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article%20view/40>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: teori tes klasik and respon [Characteristics of items: classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <https://doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan winstep* [Analysis of research data with rash and winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122->

[1?inline=1](#)

- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <http://jurnal.untad.ac.id/jurnal/index.php/JEPMT/article/view/14181>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :
Class/Department :
School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

1. Given an arithmetic sequence: 3, 8, 13, 18,
The formula for the nth term of the sequence is
A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
C. $U_n = 4n - 1$
Reason:
2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.
The number of terms in the sequence is...
A. 12 D. 23
B. 13 E. 24
C. 22
Reason:
3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
A. 5 D. 8
B. 6 E. 11
C. 7
Reason:
4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
A. 308 D. 344
B. 318 E. 354
C. 326
Reason:
5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
Then the middle term of the sequence is ...
A. 53 D. 11
B. 52 E. 10
C. 20
D. 11
E. 10
Reason:
6. Given the arithmetic sequence: 4, 10, 16, 22,
If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
A. 18 D. 24
B. 20 E. 26
C. 22
Reason:
7. The nth term of an arithmetic series is $U_n = 3n - 5$.
The formula for the sum of the first n terms of the series is ...
A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
C. $S_n = \frac{n}{2}(3n - 4)$
Reason:
8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....
A. $5n - 20$ D. $2n - 20$
B. $5n - 10$ E. $2n - 10$
C. $2n - 30$
Reason:
9. The sum of all integers between 100 and 300 which are divisible by 5 is ... $S_n = \frac{n}{2}(3n - 7)$
A. 8.200 D. 7.600
B. 8.000 E. 7.400
C. 7.800
Reason:
10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
A. 24 D. 27
B. 25 E. 28
C. 26
Reason:

11. The middle term of an arithmetic sequence is 25. If the difference is 4 and the 5th term is 21.
Then the sum of all the terms in the sequence is ...
A. 175 D. 295
B. 189 E. 375
C. 275
Reason:
12. A number of candies are distributed among five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces
A. 60 D. 75
B. 65 E. 80
C. 70
Reason:
13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...
A. 564 D. 45
B. 276 E. 36
C. 48
Reason:
14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...
A. 9 D. 12
B. 10 E. 13
C. 11
Reason:
15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...
A. 32 D. 256
B. 64 E. 512
C. 128
Reason:
16. The value of the middle term of the geometric sequence: 6, 3, ..., $3/512$ is ...
A. $\frac{1}{16}$ D. $\frac{4}{16}$
B. $\frac{1}{16}$ E. $\frac{1}{16}$
C. $\frac{3}{16}$
Reason:.....
17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm
A. 18 D. 35
B. 24 E. 40,5
C. 27,5
Reason:
18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...
A. $\frac{3}{4}$ D. $-\frac{1}{2}$
B. $\frac{1}{4}$ E. $-\frac{3}{4}$
C. $\frac{1}{3}$
Reason:
19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is.... m
A. 60 D. 90
B. 70 E. 100
C. 80
Reason:
20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.
The value of $3a + b$ is ...
A. 8 D. 14
B. 10 E. 20
C. 12
Reason:
21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$.
If $K = L$, then c is ...
A. 12 D. 15
B. 13 E. 16
C. 14
Reason:
22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$.
Then $(A + C) - (A + B)$ is ...
A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$
C. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$
Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah ...
- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 1 & 3 & 1 \\ 0 & 7 & 6 \\ 2 & 4 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 5 & 6 \\ 1 & 3 & 1 \\ 2 & 4 & 5 \end{bmatrix}$
- Reason:

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.
 If $A + B = C$, then $x + y = \dots$
 A. -5 D. 3
 B. -1 E. 5
 C. 1
- Reason:
25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah ...
- A. $\begin{bmatrix} 13 & 42 \\ 26 & 84 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \\ 30 & 36 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \\ 26 & 42 \end{bmatrix}$ E. $\begin{bmatrix} 13 & 84 \\ 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \\ 26 & 42 \end{bmatrix}$
- Reason:

26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$
 then $A(B - C) = \dots$
- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$
- Reason:
27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...
- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$
- Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix.
 Value of x that satisfies is ...
- A. -5 D. 3
 B. -4 E. 4
 C. -3
- Reason:
29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ...
- A. 0 D. 2
 B. 1 E. 4
 C. 2
- Reason:

30. Transpose matrix P adalah P^t . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^t)^{-1}$ is ...
- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$
- Reason:
31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$.
 Inverse matrix AB adalah $(AB)^{-1} = \dots$
- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$
- Reason:

32. The roots of the quadratic equation $x^2 - 3x + 4 = 0$ are ...
- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$
- Reason:
33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...
- A. -2 dan $\frac{5}{6}$ D. -2 dan $-\frac{6}{5}$
 B. 2 dan $-\frac{5}{6}$ E. -2 dan $\frac{6}{5}$
 C. 2 dan $\frac{6}{5}$
- Reason:
34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...
- A. -4 D. 2
 B. -2 E. 4
 C. 0
- Reason:

35. The roots of the quadratic equation $2x^2 - 3x - 9 = 0$ are x_1 dan x_2 . Value of $x_1^2 + x_2^2$ is ...

- A. $11\frac{1}{4}$
 B. $6\frac{3}{4}$
 C. $2\frac{1}{4}$
 D. $-6\frac{3}{4}$
 E. $-11\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...

- A. $y = x^2 - 2x + 1$
 B. $y = x^2 - 2x + 3$
 C. $y = x^2 + 2x - 1$
 D. $y = x^2 + 2x + 1$
 E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0, -4)$ then the value of $f(7)$ is ...

- A. -16
 B. -17
 C. -18
 D. -19
 E. -20

Reason:

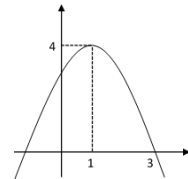
36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α dan β . The quadratic equation $(\alpha - 2)$ dan $(\beta - 2)$ is ...

- a. $x^2 + 6x + 5 = 0$
 b. $x^2 + 6x + 7 = 0$
 c. $x^2 + 6x + 11 = 0$
 d. $x^2 - 2x + 3 = 0$
 e. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation ? ...

- A. $y = x^2 + 2x + 3$
 B. $y = x^2 - 2x - 3$
 C. $y = -x^2 + 2x - 3$
 D. $y = -x^2 - 2x + 3$
 E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...

- A. $y = -x^2 + 2x - 3$
 B. $y = -x^2 + 2x + 3$
 C. $y = -x^2 - 2x + 3$
 D. $y = -x^2 - 2x - 5$
 E. $y = -x^2 - 2x + 5$

Reason:



Tanggal/Date: 22 April 2022

Formulir Kiriman Uang Remittance Application

Penerima / Beneficiary Penduduk / Resident Bukan Penduduk / Not Resident
 Nama / Name: Ahmet Emin Saray
 Alamat / Address: ...
 Kota / City: ... Negara / Country: TURKEY

Jenis Pengiriman / Type of Transfer: LLS/Clearing Draft RTGS SWIFT

Sumber Dana / Source of Fund: Tunai / Cash Cek / BG No. Debit Rek. / Debit Acc. No. 2070742750

Mata Uang / Currency: IDR USD

Bank Penerima / Beneficiary Bank: ...
 Kota / City: ... Negara / Country: TURKEY
 No. Rek./Acc. No.: ...

Jumlah / Amount	Kurs / Rate	Nilai Total / Amount
...

Pengirim / Remitter Penduduk / Resident** Bukan Penduduk / Non Resident**
 Nama / Name: Sugeng Sutiarno
 Alamat / Address: ...
 Kota / City: ... Negara / Country: Indonesia

Terbilang / Amount in Words: ...

Biaya / Charge	Nilai / Amount in Foreign Exchange	Kurs / Rate	Nilai Total / Amount
Komis / Commission			...
Pengiriman / Handling			...
Bank Koresponden / Correspondent Bank			...
Jumlah Biaya / Amount Charge			...

Biaya dari bank koresponden dibebankan ke rekening / Correspondent bank charges are for account of
 Penerima/Beneficiary Pengirim/Remitter Sharing

(Signature and Stamp)
 Pejabat Bank / Bank Officer: ...
 Penohon / Applicant: ...

001/003/WAT.L/01/000

* Transaksi non-perkulia di atas Rp. 100 juta wajib mengisi form PAMN-KYCC/Transaction by resident/amounting over Rp. 100 million must fill in the PAMN-KYCC Form
 ** Transaksi oleh bukan penduduk di atas USD 10.000 atau ekuivalennya wajib mengisi form LLD1/Transaction by non-resident/amounting over USD 10,000 or its equivalent must fill in the LLD1 Form

PT. BANK NEGARA INDONESIA (Persero), Tbk
CABANG : TANJUNG KARANG

IBCC - Maintenance (S10)

Teller ID : 00159
Date : 22/04/2022
Time : 09:29:56

Sender's Reference:
:20:810TKR00019222
Bank Operation Code:
:23B:CRED
Value Date/Currency/Interbank Settled Amount:
:32A:220422USD650,
Ordering Customer:
:50K:/0000000070742756
BPK SUGENG SUTIARSO
JL NUNYAI DALAM LKII RAJABASA
BANDAR LAMPUNG INDONESIA
Ordering Institution:
:52A:BNINIDJAMM
Account With Institutions:
:57A:FNNETRISAM
Beneficiary Customer:
:59:/TR660011100000000088177946
AHMET CEZMI SAVAC
DEGIRMICEM DISTRICT OZGURLUK STR
32B 27090 GAZIANTEP
TURKEY
Remittance Information:
:70:/SUGENG SUTIARSO
//BANDAR LAMPUNG INDONESIA
Details Of Charges:
:71A:00K
Sender to Receiver Information:
:72:/650USD
//BU JER MANUSCRIPT
///ID 21112502244011

22 APR 2022



M. RYNALDIA.
B808159


BNI

REFERENSI : 3107MM00019222

NO. TRX. : 88159 909936 96962 TRAN 22/04/2022 09:23:08
NO. REK. : 00000007074275C SUNGUNG DUTIANGO
JUMLAH : IDR 398,375 1568
142 PANJUNG KAJANG

NO. TRX. : 88159 909936 96962 TRAN 22/04/2022 09:23:08
NO. REK. : 142360420001001 PENDAPATAN PROPISI KU
JUMLAH : IDR 36,000 1568
142 TANJUNG KARANG

NO. TRX. : 88159 909936 96962 TRAN 22/04/2022 09:23:08
NO. REK. : 1423604602010001 Pendapatan Restitusi B
JUMLAH : IDR 363,375 1568
142 TANJUNG KARANG

NO. TRX. : 88159 909936 96962 TRAN 22/04/2022 09:23:08
NO. REK. : 00000007074275C SUNGUNG DUTIANGO
JUMLAH : IDR 9,447,750 1568
142 TANJUNG KARANG

NO. TRX. : 88159 909936 96962 TRAN 22/04/2022 09:23:08
NO. REK. : 142840200101001 MU YAKIR
JUMLAH : USD 650 1568
142 TANJUNG KARANG

22 APR 2022



M. RYNALDIA.
BB08159


BNI



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

**Received your payment & asking the copyright transfer agreement
(ID#21112502244011)**

3 messages

European Journal of Educational Research <editor@eu-jer.com>

Fri, Apr 22, 2022 at 5:24 PM

Reply-To: European Journal of Educational Research <editor@eu-jer.com>

To: European Journal of Educational Research <sugeng.sutiarso@fkip.unila.ac.id>

Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Dear Dr. Sugeng Sutiarso,

We have received your payment about your paper entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" ID#21112502244011. Thanks.

We kindly ask from you to sign the copyright transfer agreement for your paper. After all author(s) signed, please scan and send via email to me **as soon as possible**. Please download the pdf file of this agreement from this link : <https://eu-jer.com/EU-JER-copyright-transfer-agreement.pdf> You can use e-signature, if you have. Also you can use your mobil phone as a scanner. If the other author live in another city, he/she sign the paper and send this paper via email. Than you can sign on this paper.

We are preparing the galley proof of your paper. We will send it to you in order to check before publication. The preparing of galley proofs may take some time because of our intensity. Thank you for your patience.

We are looking forward to getting copyright transfer agreement.

Best regards,

Ahmet C. Savas, Ph.D.

Editor, European Journal of Educational Research

editor@eu-jer.comwww.eu-jer.com

SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Sat, Apr 23, 2022 at 8:36 AM

To: European Journal of Educational Research <editor@eu-jer.com>

Dear Ahmet C. Savas, Ph.D.

Editor, European Journal of Educational Research

Here, I attach the copyright transfer agreement.

Best regards,

Sugeng Sutiarso

Lampung University

[Quoted text hidden]

 **EU-JER-copyright-transfer-agreement_Sugeng Sutiarso.pdf**
1110K

Editor - European Journal of Educational Research <editor@eu-jer.com>

Sat, Apr 23, 2022 at 2:56 PM

To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Thanks, received.

Best regards,

Ahmet C. Savas, Ph.D.

Editor, European Journal of Educational Research

Copyright Transfer Agreement

European Journal of Educational Research [EU-JER] ("the Proprietor") will be pleased to publish your article ("the Work"), tentatively entitled

Developing Assessment Instrument Using Polytomous Response in Mathematics

in the *EU-JER* ("the Journal") if the Work is accepted for publication. The undersigned authors transfer all copyright ownership in and relating to the Work, in all forms and media, to the Proprietor in the event that the Work is published. However, this agreement will be null and void if the Work is not published in the Journal.

The undersigned authors warrant that the Work is original, is not under consideration by another journal, and has not been previously published.

(This agreement must be signed by all authors. A photocopy of this form may be used if there are more than 10 authors.)



Sugeng Sutiarmo

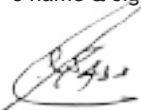
April 23, 2022

Author's name & signature

Date

Author's name & signature

Date



Undang Rosidin

April 23, 2022

Author's name & signature

Date

Author's name & signature

Date



Aan Sulistiawan

April 23, 2022

Author's name & signature

Date

Author's name & signature

Date

Author's name & signature

Date

Author's name & signature

Date

Author's name & signature

Date

Author's name & signature

Date

7/20/22, 1:14 PM

unila.ac.id Mail - Received your payment & asking the copyright transfer agreement (ID#21112502244011)

editor@eu-jer.com
www.eu-jer.com

[Quoted text hidden]



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Proofreading (ID#21112502244011)

3 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Fri, May 6, 2022 at 7:38 PM

Dear Dr. Sutiarso,

Thank you for your kind email. Please see the attached file as the proofreading of your paper.

We are preparing the galley proof of your paper.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 4/23/2022 4:36 AM, SUGENG SUTJARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

Here, I attach the copyright transfer agreement.

Best regards,
Sugeng Sutiarso
Lampung UniversityOn Fri, Apr 22, 2022 at 5:24 PM European Journal of Educational Research <editor@eu-jer.com>
wrote:

Dear Dr. Sugeng Sutiarso,

We have received your payment about your paper entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" ID#21112502244011. Thanks.

We kindly ask from you to sign the copyright transfer agreement for your paper. After all author(s) signed, please scan and send via email to me **as soon as possible**. Please download the pdf file of this agreement from this link : <https://eu-jer.com/EU-JER-copyright-transfer-agreement.pdf> You can use e-signature, if you have. Also you can use your mobil phone as a scanner. If the other author live in another city, he/she sign the paper and send this paper via email. Than you can sign on this paper.

We are preparing the galley proof of your paper. We will send it to you in order to check before publication. The preparing of galley proofs may take some time because of our intensity. Thank you for your patience.

We are looking forward to getting copyright transfer agreement.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com**SUTJARSO - PROOFREADING.docx**

1706K

Developing Assessment Instrument Using Polytomous Response in Mathematics

RH: Developing Assessment Instrument Using Polytomous

Sugeng Sutiarto^{1*}, University of Lampung, Indonesia, <https://orcid.org/0000-0003-4097-6000>, sugeng.sutiarto@fkip.unila.ac.id

Undang Rosidin, University of Lampung, Indonesia, <https://orcid.org/0000-0003-1589-2403>, undang.rosidin@fkip.unila.ac.id

Aan Sulistiawan, Vocational School, Indonesia, aansulistiawan95@guru.smp.belajar.id

* Corresponding Author

University of Lampung. Sumantri Brojonegoro No. 1, Bandar Lampung, Indonesia.

Authorship Contribution Statement

Sutiarto: Conceptualization, design, development of instruments, analysis, article writing, final approval. Rosidin: Design, development of instruments, analysis, editing. Sulistiawan: development of instruments, collect data, analysis, editing.

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used an open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) the open polytomous response test has a good category according to classical and modern theory. However, the discrimination power of test items in classical theory needs several revisions, (2) the assessment instrument using the polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments

when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools was divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).

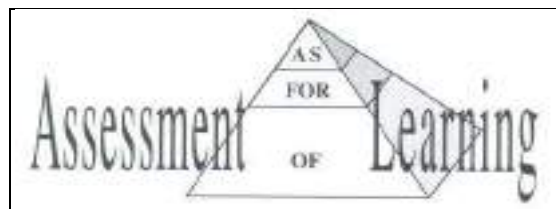


Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with specific rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over

essay test are firstly, the test can be conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weaknesses. The multiple-choice test is not able to portray the actual abilities of students, the answers of this test tend to be guessing games or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). A multiple-choice test with only two answer choices is called a "dichotomous test," and a multiple-choice test with more than two answer choices is called a "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or henceforth called as polytomous response test (Suwanto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as a test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), a test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material

(Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into an open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response tests for students in college and senior or junior school. Students in college and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools is not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomus test. Often, students pay less attention during math exams for several

reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmawati, 2020). Therefore, the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomous response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing the item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because, in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows:(1) Does the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory?and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's (2013) model, with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

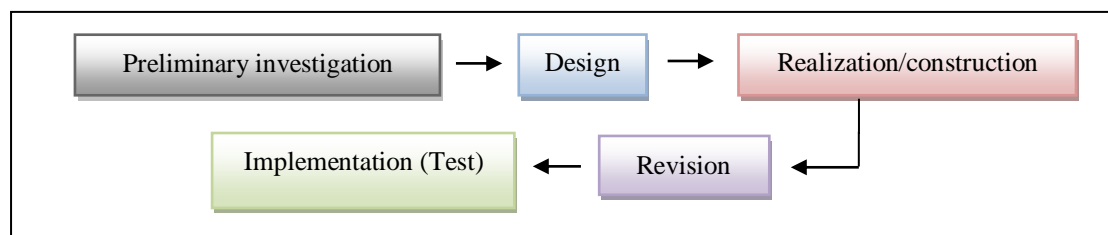


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the

items with the indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. *Questionnaire data analysis (qualitative analysis)*

There are two sets of questionnaire data namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the

two questionnaires were analyzed descriptively. Specifically for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. *Test data analysis (empirical analysis)*

After conducting the content validity test, the researchers conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to have good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program was used for classical theory and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze

polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep) because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Correlation column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit's MNSQ value is 0.5 to 1.5 and the Pt-Measure Correlation is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay tests and 20% of teachers used multiple-choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of

teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

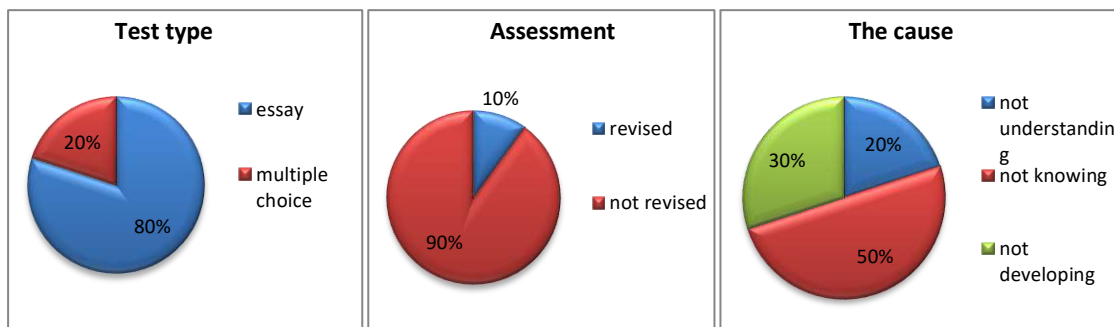


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD

format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly

calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised
4	0.540	good	0.304	Good	24	0.438	good	-0.071	Revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	Revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	Revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	Revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	Revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	Revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	Revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	Revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	Revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	Revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	Revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	Revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	Revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	Revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	Revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	Revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

Analysis of Test Data with Modern Theory

The Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first-factor Eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first-factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

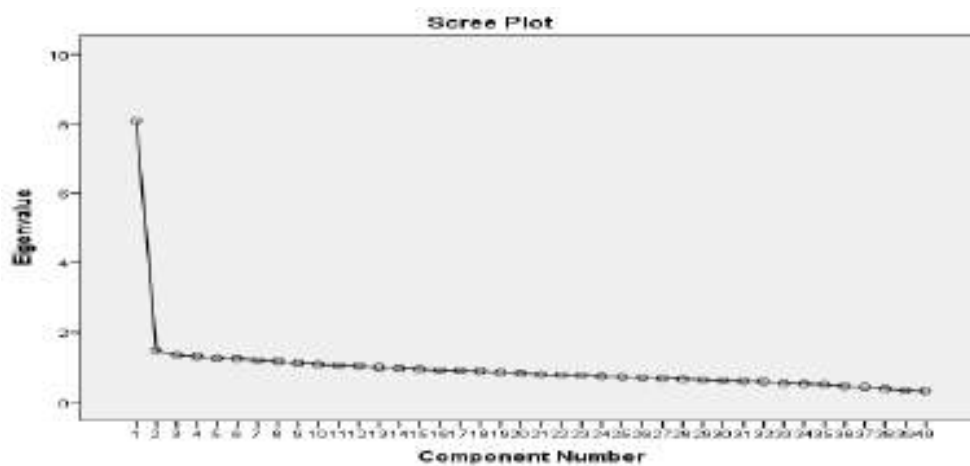


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The

results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	<i>K1</i>	<i>K2</i>	<i>K3</i>	<i>K4</i>	<i>K5</i>	<i>K6</i>	<i>K7</i>	<i>K8</i>	<i>K9</i>	<i>K10</i>
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-MeasureCorrelation is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage,

the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.43	.07	.97	-4	.98	-5	.73	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	55.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-5	.97	-5	.43	.44	48.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	964	413	-.02	.07	.97	-4	.98	-4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.3	1.10	1.6	.49	.44	42.8	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-7	.96	-6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-4	.97	-4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.3	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.1	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-4	.97	-4	.31	.44	50.0	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-5	.97	-5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.25	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	881	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-6	.96	-6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.5	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.69	-6.3	.66	-6.2	.39	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained were presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

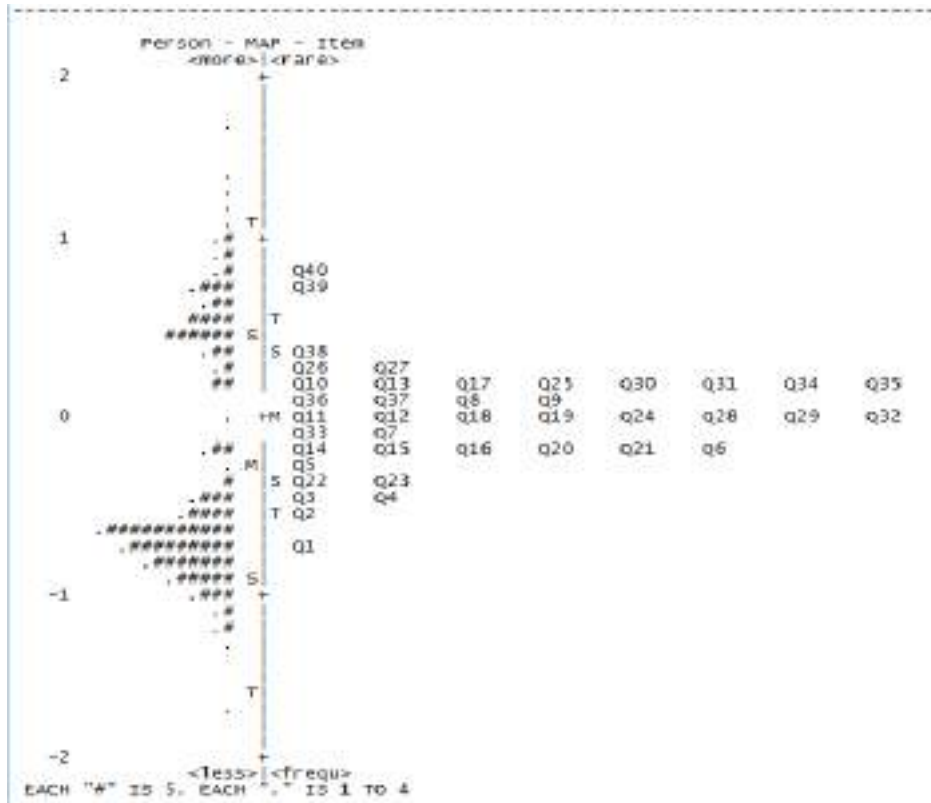


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (PT-Measure Correlation column). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

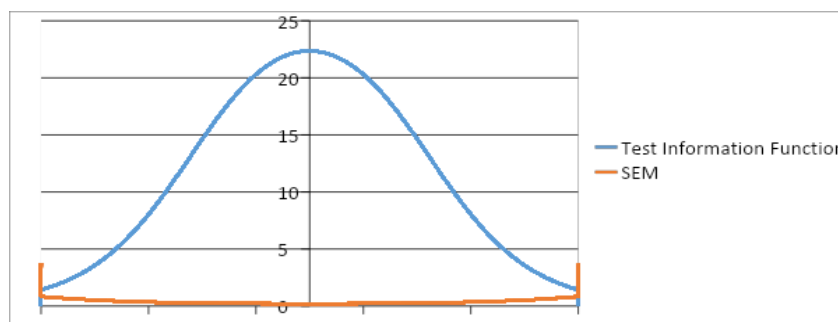


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in the vocational school was based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

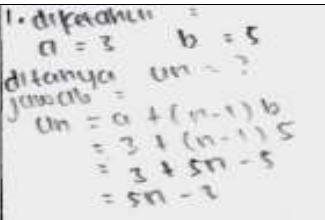
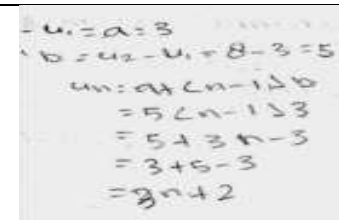
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, ... The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	 <p>1. diketahui $a = 3$ $b = 5$ ditanya $U_n = ?$ jawab $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	 <p>- $u_1 = a = 3$ $b = u_2 - u_1 = 8 - 3 = 5$ $u_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term) and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

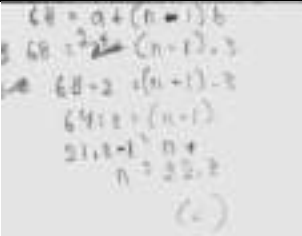
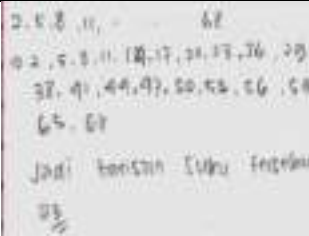
Question2:	Pattern 1	Pattern 2
<p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>		

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

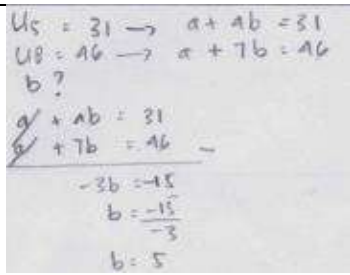
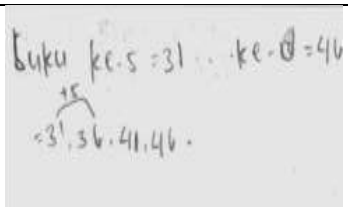
Question 3:	Pattern 1	Pattern 2
<p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>		

Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the n th term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the n th term.

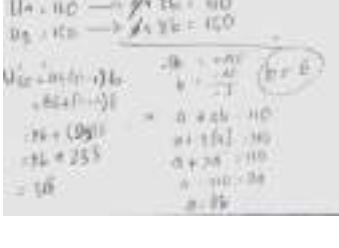
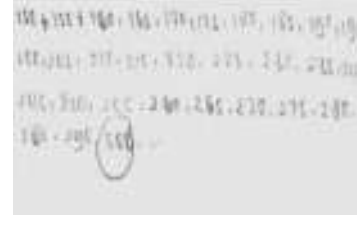
Question4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>	 <p>Handwritten student work for Pattern 1. It shows the general formula for an arithmetic sequence: $U_n = a + (n-1)b$. The student substitutes $U_4 = 110$ and $U_9 = 150$ to form a system of equations: $a + 3b = 110$ and $a + 8b = 150$. They subtract the first equation from the second to get $5b = 40$, leading to $b = 8$. Then they substitute $b = 8$ back into the first equation to get $a = 106$. Finally, they calculate $U_{30} = 106 + (30-1) \cdot 8 = 318$.</p>	 <p>Handwritten student work for Pattern 2. The student lists terms of the sequence starting from the 4th term: 110, 118, 126, 134, 142, 150, 158, 166, 174, 182, 190, 198, 206, 214, 222, 230, 238, 246, 254, 262, 270, 278, 286, 294, 302, 310, 318. The 30th term, 318, is circled.</p>

Figure 13. An Example of Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

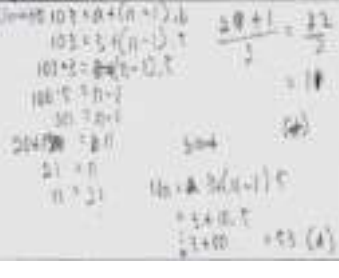

Question 5:	Answer Pattern 1	Answer Pattern 2
<p>An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...308</p> <p>A. 53 B. 52 C. 20 D. 11 E. 10</p> <p>Reason:</p>	 <p>Handwritten student work for Answer Pattern 1. The student uses the formula for the nth term: $U_n = a + (n-1)b$. They substitute $U_1 = 3$ and $U_n = 103$ to get $103 = 3 + (n-1)b$. They also use the formula for the sum of an arithmetic sequence: $S_n = \frac{n}{2}(2a + (n-1)b)$. They substitute $S_n = 308$ and $a = 3$ to get $308 = \frac{n}{2}(2 \cdot 3 + (n-1)b)$. They solve for n and find $n = 20$.</p>	 <p>Handwritten student work for Answer Pattern 2. The student lists terms of the sequence: 3, 8, 13, 18, 23, 28, 33, 38, 43, 48, 53, 58, 63, 68, 73, 78, 83, 88, 93, 98, 103. The middle term, 53, is circled.</p>

Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test

and an essay. Multiple choice test is easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score. Then, the result is multiplied by 10 to get a value range of 0–10 or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through

the open polytomus response test is more secondhand and complete than other assessment instruments, such as multiple-choice tests(Gierl et al., 2017) or essay tests(Putri et al., 2020).

Discussion

This research is development research aimed at developing a good mathematical assessment instrument using polytomous responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of the classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to the modern theory before revising or replacing the item.

Research on learning assessment with open response polytomous was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

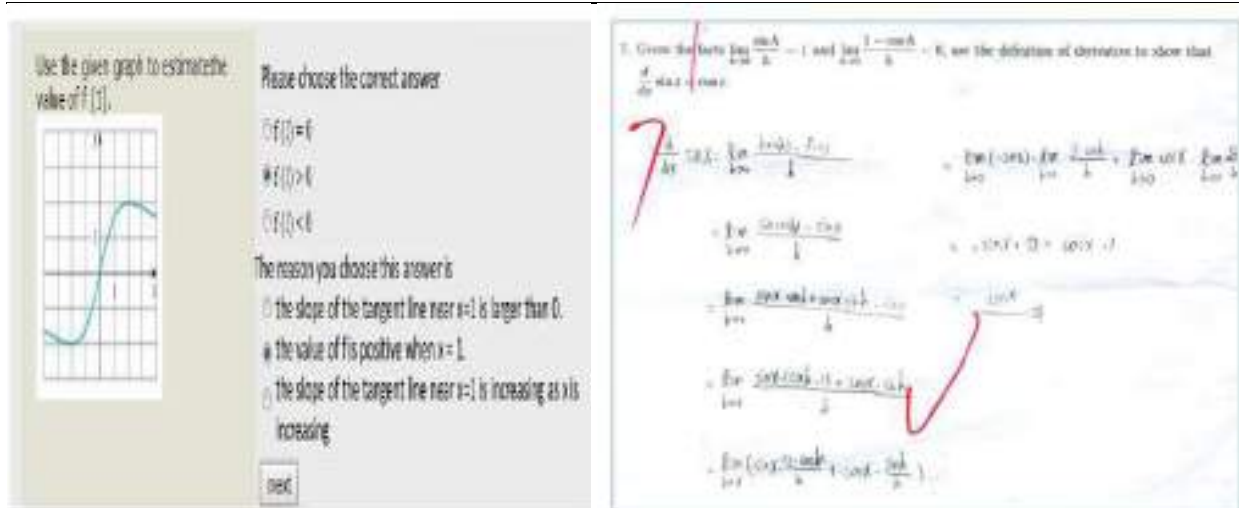


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous Response Test

Another study on learning assessment with open response polychromous was conducted by Ayanwale(2021).Ayanwale's research compares two polytomous response test analysis methods namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

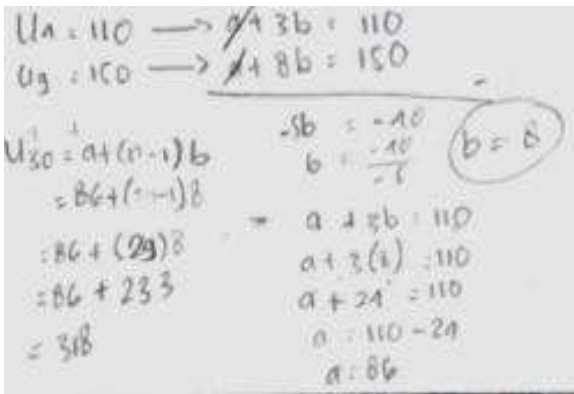
Other studies related to classical and modern theory were conducted by Sarea (2018) and Saepuzaman et al.(2021). Sarea's research states that the response polytomous test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomous test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

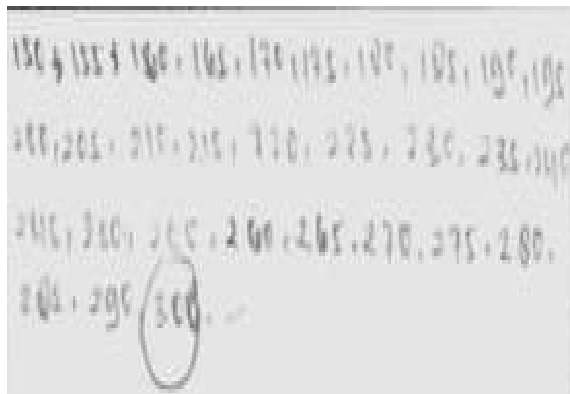
The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

A. 308
B. 318
C. 326
D. 344
E. 354

Reason:



(i)



(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, using a drawing or model, analogy, and formula. Syahlan(2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syahlan's opinion,

students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. Thus, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, it is suggested to conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2),95-113. <https://journal.unhas.ac.id/index.php/jmsk/article/view/3402/1939>

- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *JurnalSuluh Pendidikan*, 17(1), 32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan*[Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://ejournal.upsi.edu.my/index.php/EJSMT/article/view/5029>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. [https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_\(1\)%20Lorna.pdf](https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_(1)%20Lorna.pdf)
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geofrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. ResearchGate. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of*

- Educational Research*, 87(6), 1082–1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: Challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar [Equalization of the dichotomous and polytomus mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how? In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. http://seminar.uny.ac.id/icriems/sites/seminar.uny.ac.id/icriems/files/proceeding2018/ME69_Asma%20Khiyarunnisa.pdf
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148.

<https://usnsj.com/index.php/JME/article/view/1607>

- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset pemasaran*[Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://jurnal.uns.ac.id/ijscs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah*[Standards of content for primary and secondary education]. Indonesian Government publicationservice. [https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud Tahun2016 Nomor021.pdf](https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud_Tahun2016_Nomor021.pdf)
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://journal.unnes.ac.id/sju/index.php/jere/article/view/46133>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://journal.unnes.ac.id/sju/index.php/ujme/article/view/12643>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*[Item response theory and its application]. NuhaMedika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran*[Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163.

<http://dx.doi.org/10.30587/postulat.v1i2.2094>

- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst: Journal of Physics Education and Its Application/Karst: Jurnal Pendidikan Fisika dan Terapannya*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article%20/view/40>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: teori tes klasik and respon [Characteristics of items: classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <https://doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan Winstep* [Analysis of research data with rash and Winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122->

[1?inline=1](#)

- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <http://jurnal.untad.ac.id/jurnal/index.php/JEPMT/article/view/14181>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian*[Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student:
Class/Department :
School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

- | | |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>1. Given an arithmetic sequence: 3, 8, 13, 18,
The formula for the nth term of the sequence is
A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
C. $U_n = 4n - 1$</p> <p>Reason:</p> | <p>2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.
The number of terms in the sequence is...
A. 12 D. 23
B. 13 E. 24
C. 22</p> <p>Reason:</p> |
| <p>3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
A. 5 D. 8
B. 6 E. 11
C. 7</p> <p>Reason:</p> | <p>4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
A. 308 D. 344
B. 318 E. 354
C. 326</p> <p>Reason:</p> |
| <p>5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
Then the middle term of the sequence is ...
A. 53 D. 11
B. 52 E. 10
C. 20
D. 11
E. 10</p> <p>Reason:</p> | <p>6. Given the arithmetic sequence: 4, 10, 16, 22,
If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
A. 18 D. 24
B. 20 E. 26
C. 22</p> <p>Reason:</p> |
| <p>7. The nth term of an arithmetic series is $U_n = 3n - 5$.
The formula for the sum of the first n terms of the series is ...
A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
C. $S_n = \frac{n}{2}(3n - 4)$</p> <p>Reason:</p> | <p>8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....
A. $5n - 20$ D. $2n - 20$
B. $5n - 10$ E. $2n - 10$
C. $2n - 30$</p> <p>Reason:</p> |
| <p>9. The sum of all integers between 100 and 300 which are divisible by 5 is ... $S_n = \frac{n}{2}(3n - 7)$
A. 8.200 D. 7.600
B. 8.000 E. 7.400
C. 7.800</p> <p>Reason:</p> | <p>10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
A. 24 D. 27
B. 25 E. 28
C. 26</p> <p>Reason:</p> |
| <p>11. The middle term of an arithmetic sequence is 25. If the difference is 4 and the 5th term is 21.</p> | <p>12. A number of candies are distributed among five children according to the rules of an arithmetic</p> |

Then the sum of all the terms in the sequence is ...

- A. 175D.295
- B. 189E.375
- C. 275

Reason:

sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces

- A. 60D.75
- B. 65E.80
- C. 70

Reason:

13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...

- A. 564D.45
- B. 276E.36
- C. 48

Reason:

14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...

- A. 9D.12
- B. 10E.13
- C. 11

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...

- A. 32D.256
- B. 64E.512
- C. 128

Reason:

16. The value of the middle term of the geometric sequence: 6, 3, ..., $3/512$ is ...

- A. $\frac{1}{16}$ D. $\frac{4}{16}$
- B. $\frac{2}{16}$ E. $\frac{5}{16}$
- C. $\frac{3}{16}$

Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm

- A. 18D.35
- B. 24E.40,5
- C. 27,5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...

- A. $\frac{3}{4}$ D. $-\frac{1}{2}$
- B. $-\frac{1}{4}$ E. $-\frac{3}{4}$
- C. $\frac{1}{3}$

Reason:

19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is...m

- A. 60D.90
- B. 70E.100
- C. 80

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.

The value of $3a + b$ is ...

- A. 8D.14
- B. 10E.20
- C. 12

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$.

If $K = L$, then cis...

- A. 12D.15
- B. 13E.16
- C. 14

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$.

Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$
- C. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah ...

- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah ...

- A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...

- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix

- determinant A is ...
 A. 0 D. 2
 B. 1 E. 4
 C. 2

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$.

Inverse matrix AB adalah $(AB)^{-1} = \dots$

- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$

Reason:

33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...

- A. -2 dan $\frac{5}{6}$ D. -2 dan $-\frac{6}{5}$
 B. 2 dan $-\frac{5}{6}$ E. -2 dan $\frac{6}{5}$
 C. 2 dan $\frac{6}{5}$

Reason:

35. The roots of the quadratic equation $2x^2 - 3x - 9 = 0$ are x_1 dan x_2 . Value of $x_1^2 + x_2^2$ is ...

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.

If $A + B = C$, then $x + y = \dots$

- A. -5 D. 3
 B. -1 E. 5
 C. 1

Reason:

26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$

then $A(B - C) = \dots$

- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix.

Value of x that satisfies is

- A. -5 D. 3
 B. -4 E. 4
 C. -3

Reason:

30. Transpose matrix P adalah P^t . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^t)^{-1}$ is ...

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$

Reason:

32. The roots of the quadratic equation $3x^2 - 4x - 4$ are ...

- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$

Reason:

34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...

- A. -4 D. 2
 B. -2 E. 4
 C. 0

Reason:

36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α dan β . The quadratic

A. $11\frac{1}{4}$ D. $-6\frac{3}{4}$

B. $6\frac{3}{4}$ E. $-11\frac{1}{4}$

C. $2\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...

- A. $y = x^2 - 2x + 1$
- B. $y = x^2 - 2x + 3$
- C. $y = x^2 + 2x - 1$
- D. $y = x^2 + 2x + 1$
- E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0, -4)$ then the value of $f(7)$ is ...

- A. -16 D. -19
- B. -17 E. -20
- C. -18

Reason:

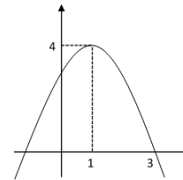
equation $(\alpha - 2)$ dan $(\beta - 2)$ is ...

- a. $x^2 + 6x + 5 = 0$
- b. $x^2 + 6x + 7 = 0$
- c. $x^2 + 6x + 11 = 0$
- d. $x^2 - 2x + 3 = 0$
- e. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation ? ...

- A. $y = x^2 + 2x + 3$
- B. $y = x^2 - 2x - 3$
- C. $y = -x^2 + 2x - 3$
- D. $y = -x^2 - 2x + 3$
- E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...

- A. $y = -x^2 + 2x - 3$ D. $y = -x^2 - 2x - 5$
- B. $y = -x^2 + 2x + 3$ E. $y = -x^2 - 2x + 5$
- C. $y = -x^2 - 2x + 3$

Reason:

SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Tue, May 10, 2022 at 5:24 PM

To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

Here I attach a revision of your proofreading suggestion.

Best regards,
Sugeng Sutiarso
Lampung University

[Quoted text hidden]

 **Ok_SUTJARSO - PROOFREADING.docx**
1702K

Editor - European Journal of Educational Research <editor@eu-jer.com>

Tue, May 10, 2022 at 7:15 PM

To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Dear Dr. Sutiarso,

Thank you for your kind email.

Your paper with proofreading corrections has been sent to galley proof service.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

[Quoted text hidden]

Developing Assessment Instrument Using Polytomous Response in Mathematics

RH: Developing Assessment Instrument Using Polytomous

Sugeng Sutiarmo^{1*}, University of Lampung, Indonesia, <https://orcid.org/0000-0003-4097-6000>, sugeng.sutiarmo@fkip.unila.ac.id

Undang Rosidin, University of Lampung, Indonesia, <https://orcid.org/0000-0003-1589-2403>, undang.rosidin@fkip.unila.ac.id

Aan Sulistiawan, Vocational School, Indonesia, aansulistiawan95@guru.smp.belajar.id

* Corresponding Author

University of Lampung. Sumantri Brojonegoro No. 1, Bandar Lampung, Indonesia.

Authorship Contribution Statement

Sutiarmo: Conceptualization, design, development of instruments, analysis, article writing, final approval. Rosidin: Design, development of instruments, analysis, editing. Sulistiawan: development of instruments, collect data, analysis, editing.

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used an open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) the open polytomous response test has a good category according to classical and modern theory. However, the discrimination power of test items in classical theory needs several revisions, (2) the assessment instrument using the polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments when giving reasons

for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools was divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).

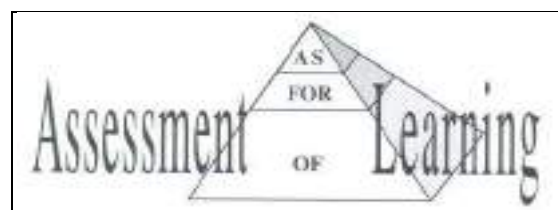


Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with specific rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over essay

test are firstly, the test can be conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weaknesses. The multiple-choice test is not able to the actual abilities of students, the answers of this test tend to be guessing games or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). A multiple-choice test with only two answer choices is called a "dichotomous test," and a multiple-choice test with more than two answer choices is called a "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or hence forth called as polytomous response test (Suwanto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as a test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), a test on calculus material (Khiyarunnisa & Retnawati, 2018), test order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021).

Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into an open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response tests for students in college and senior or junior school. Students in college and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools is not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomus test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a

complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmawati, 2020). Therefore, the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomous response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing the item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because, in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The

research problems are stated as follows: (1) Does the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory? and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's (2013) model, with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

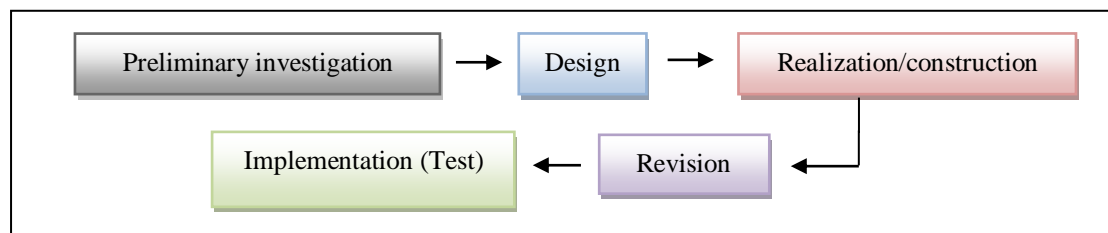


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the

items with the indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. *Questionnaire data analysis (qualitative analysis)*

There are two sets of questionnaire data namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the

two questionnaires were analyzed descriptively. Specifically for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. *Test data analysis (empirical analysis)*

After conducting the content validity test, the researchers conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to having good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the IteMan program was used for classical theory and the Winsteps program for modern theory (Sarea & Ruslan,

2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep) because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker &

Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Correlation column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit's MNSQ value is 0.5 to 1.5 and the Pt-Measure Correlation is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay tests and 20% of teachers used multiple-choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

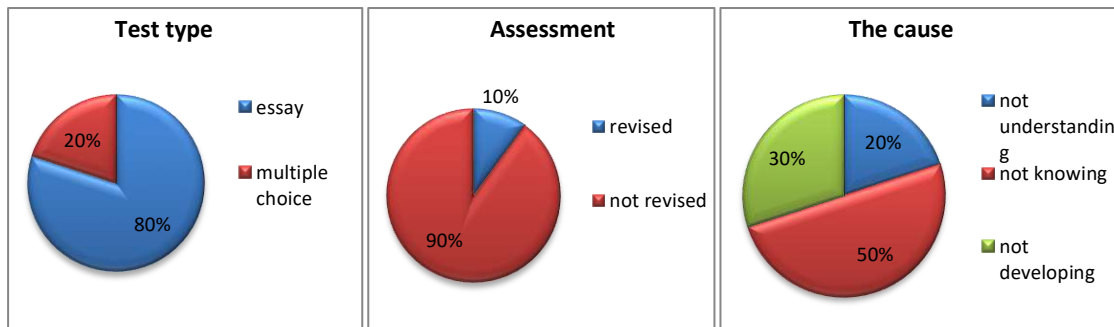


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	good	0.196	Revised	21	0.482	good	0.143	revised
2	0.487	good	0.179	Revised	22	0.535	good	0.429	good
3	0.528	good	0.214	Revised	23	0.492	good	0.250	revised
4	0.540	good	0.304	Good	24	0.438	good	-0.071	Revised
5	0.489	good	0.089	Revised	25	0.436	good	-0.107	Revised
6	0.446	good	-0.161	Revised	26	0.385	good	-0.286	Revised
7	0.438	good	-0.232	Revised	27	0.383	good	-0.321	Revised
8	0.453	good	-0.143	Revised	28	0.416	good	-0.143	Revised
9	0.414	good	-0.143	Revised	29	0.458	good	-0.125	Revised
10	0.409	good	-0.339	Revised	30	0.385	good	-0.375	Revised
11	0.438	good	-0.143	Revised	31	0.404	good	-0.321	Revised
12	0.436	good	-0.036	Revised	32	0.433	good	-0.250	Revised
13	0.400	good	-0.321	Revised	33	0.441	good	0,036	Revised
14	0.450	good	-0.036	Revised	34	0.424	good	-0.268	Revised
15	0.462	good	0.250	Revised	35	0.412	good	-0.321	Revised
16	0.453	good	-0.089	Revised	36	0.431	good	-0.304	Revised
17	0.416	good	-0.143	Revised	37	0.404	good	-0.232	Revised
18	0.419	good	-0.196	Revised	38	0.363	good	-0.482	Revised
19	0.431	good	-0.232	Revised	39	0.230	good	-0.929	Revised
20	0.441	good	-0.089	Revised	40	0.211	good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

Analysis of Test Data with Modern Theory

The Unidimensional Assumption Test

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then

used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first-factor Eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first-factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

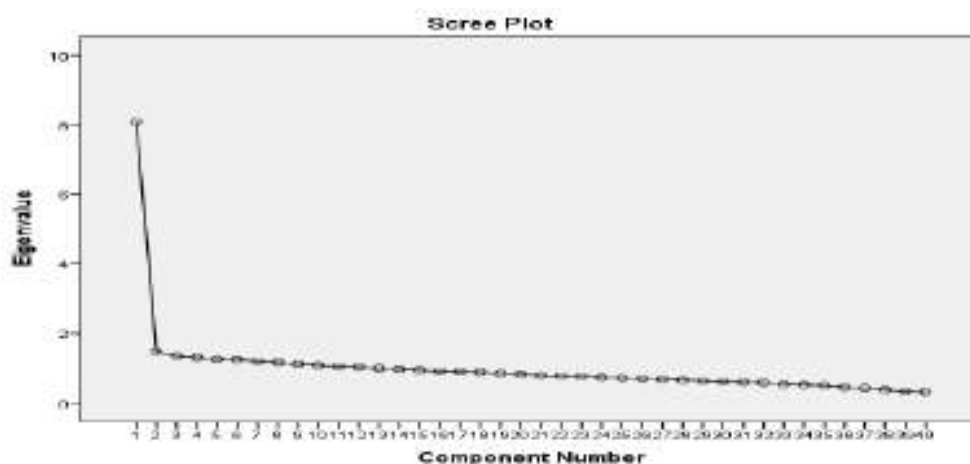


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	<i>K1</i>	<i>K2</i>	<i>K3</i>	<i>K4</i>	<i>K5</i>	<i>K6</i>	<i>K7</i>	<i>K8</i>	<i>K9</i>	<i>K10</i>
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model". If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Correlation is positive, then it can be said that the item fits the

model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		ENACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.43	.07	.97	-.4	.98	-.5	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.3	.97	-.5	.43	.44	48.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	964	413	-.03	.07	.97	-.4	.98	-.4	.39	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	-.2	1.01	-.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.1	1.10	1.1	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	-.7	1.05	-.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.3	1.08	1.3	.56	.44	43.6	50.8	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.6	50.9	Q26
27	936	413	.20	.07	1.02	-.3	1.02	-.4	.32	.44	48.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.28	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.89	-8.3	.88	-8.2	.34	.43	58.1	48.8	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.43	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained were presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

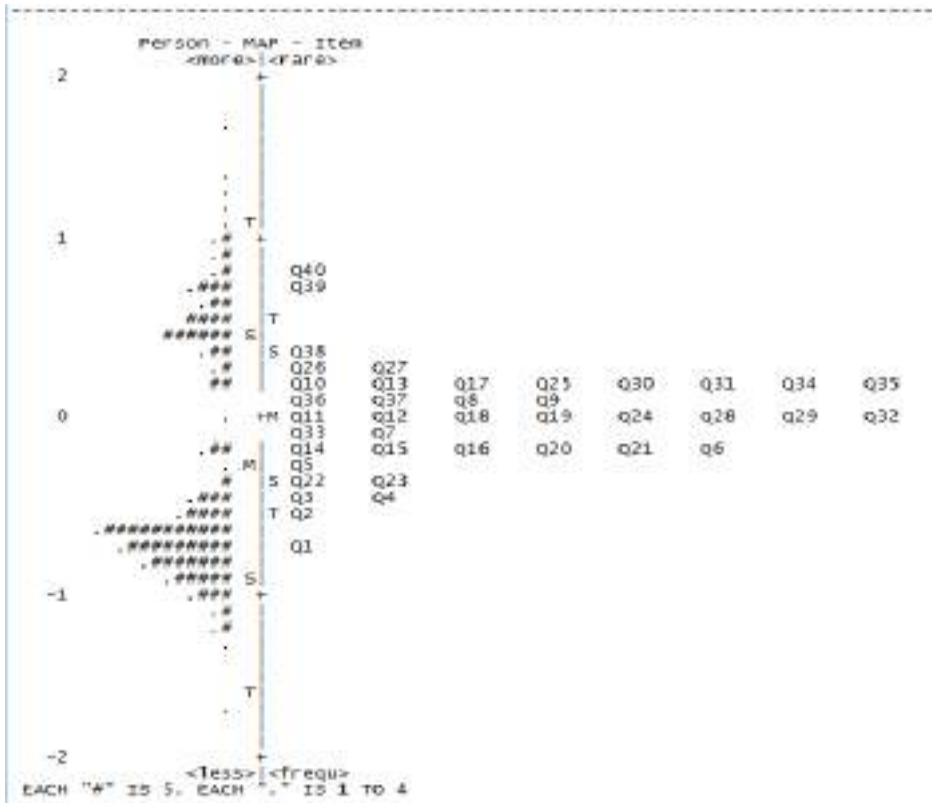


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (PT-Measure Correlation column). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

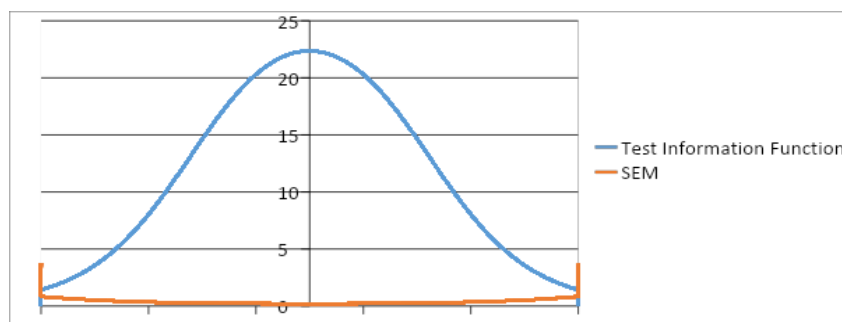


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in the vocational school was based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

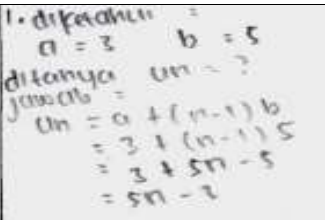
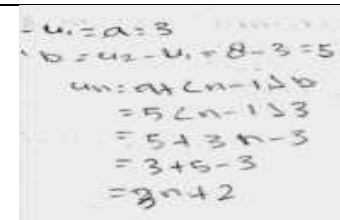
Question 1:	Pattern 1:	Pattern 2:
<p>Given an arithmetic sequence: 3, 8, 13, 18, ... The formula for the nth term of the sequence is ...</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	 <p>1. diketahui $a = 3$ $b = 5$ ditanya $U_n = ?$ jawab $U_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>	 <p>- $u_1 = a = 3$ $b = u_2 - u_1 = 8 - 3 = 5$ $u_n = a + (n-1)b$ $= 3 + (n-1)5$ $= 3 + 5n - 5$ $= 5n - 2$</p>

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term) and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

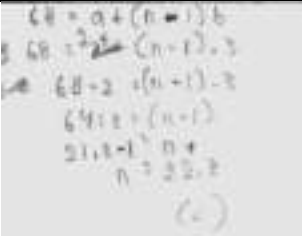
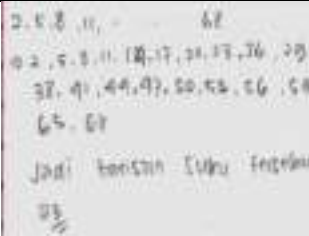
Question2:	Pattern 1	Pattern 2
<p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>	 <p> $U_n = a + (n-1)b$ $68 = 2 + (n-1) \cdot 3$ $68 - 2 = (n-1) \cdot 3$ $66 = 3(n-1)$ $22 = n-1$ $n = 22 + 1$ $n = 23$ </p>	 <p> 2, 5, 8, 11, ..., 68 02, 05, 08, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 59, 62, 65, 68 Jadi konstan (uku) frekuensi </p>

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

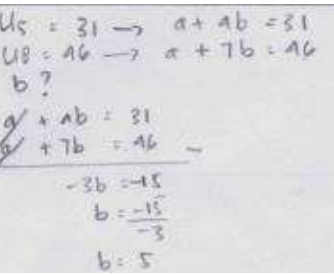
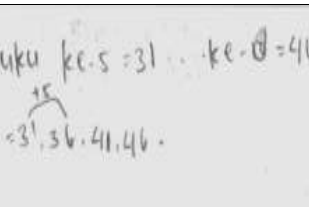
Question 3:	Pattern 1	Pattern 2
<p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>	 <p> $U_5 = 31 \rightarrow a + 4b = 31$ $U_8 = 46 \rightarrow a + 7b = 46$ $b = ?$ $a + 4b = 31$ $a + 7b = 46$ $-3b = -15$ $b = \frac{-15}{-3}$ $b = 5$ </p>	 <p> buku ke-5 = 31 .. ke-8 = 46 \uparrow $+5$ = 31, 36, 41, 46. </p>

Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

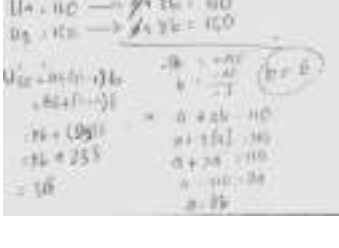
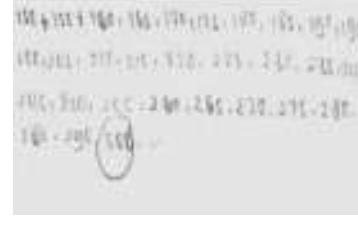
Question4:	Pattern 1	Pattern 2
<p>The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 B. 318 C. 326 D. 344 E. 354</p> <p>Reason:</p>		

Figure 13. An Example of Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

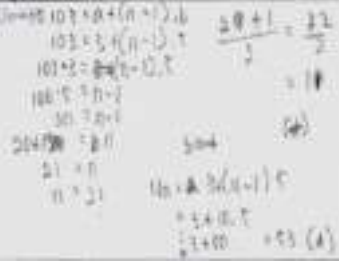

Question 5:	Answer Pattern 1	Answer Pattern 2
<p>An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...308</p> <p>A. 53 B. 52 C. 20 D. 11 E. 10</p> <p>Reason:</p>		

Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test

and an essay. Multiple choice test is easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score. Then, the result is multiplied by 10 to get a value range of 0–10 or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomus response test is more secondhand and complete than other assessment

instruments, such as multiple-choice tests (Gierl et al., 2017) or essay tests (Putri et al., 2020).

Discussion

This research is development research aimed at developing a good mathematical assessment instrument using polytomous responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of the classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to the modern theory before revising or replacing the item.

Research on learning assessment with open response polytomous was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

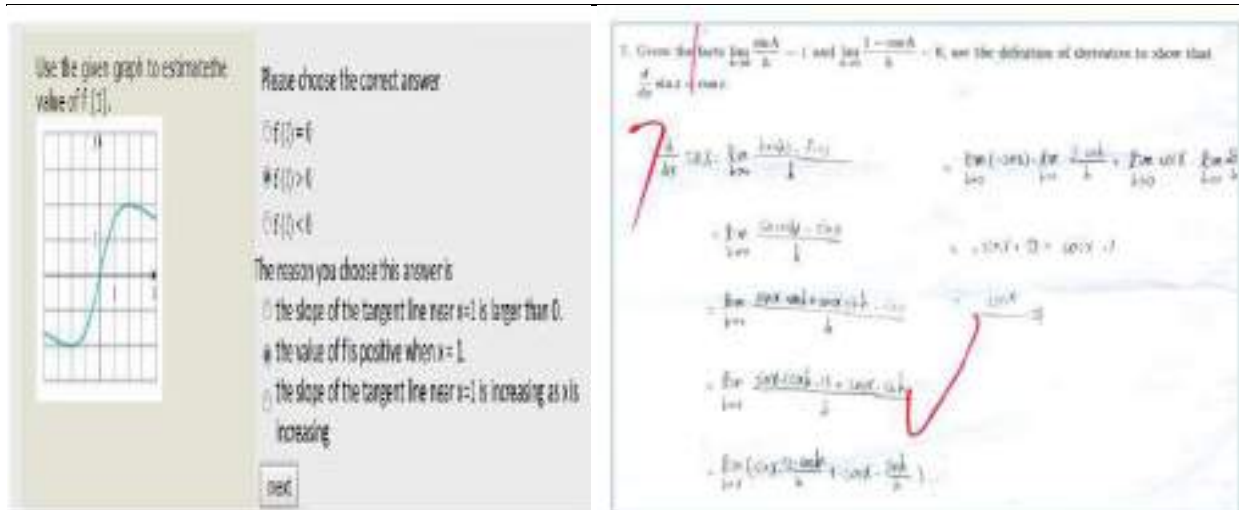


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous Response Test

Another study on learning assessment with open response polytomous was conducted by Ayanwale (2021). Ayanwale's research compares two polytomous response test analysis methods namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

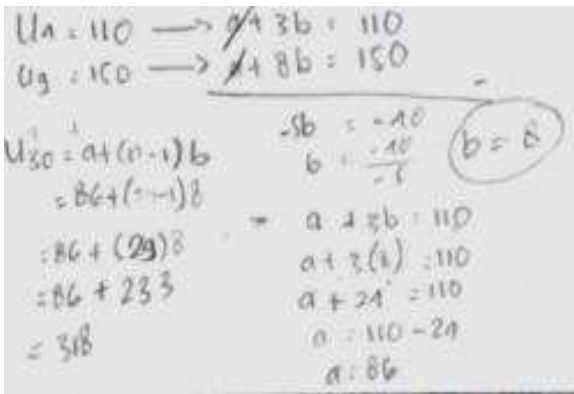
Other studies related to classical and modern theory were conducted by Sarea (2018) and Saepuzaman et al. (2021). Sarea's research states that the response polytomous test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomous test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

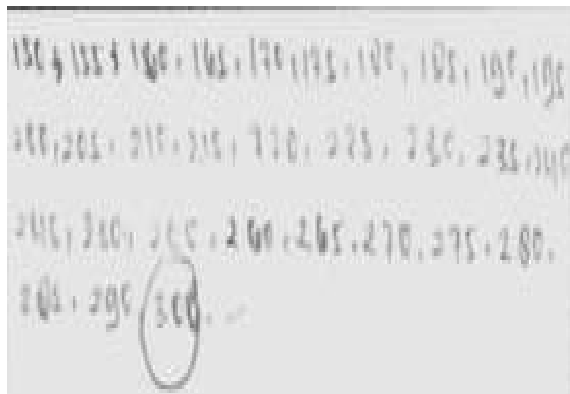
The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

A. 308
B. 318
C. 326
D. 344
E. 354

Reason:



(i)



(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, using a drawing or model, analogy, and formula. Syahlan (2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syahlan's opinion,

students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. Thus, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, it is suggested to conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2),95-113. <https://journal.unhas.ac.id/index.php/jmsk/article/view/3402/1939>

- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *JurnalSuluh Pendidikan*, 17(1), 32-37. <https://jurnal.ikipsaraswati.ac.id/index.php/suluh-pendidikan/article/view/57>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan*[Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://ejournal.upsi.edu.my/index.php/EJSMT/article/view/5029>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://iopscience.iop.org/article/10.1088/1742-6596/1155/1/012078/pdf>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. [https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_\(1\)%20Lorna.pdf](https://wlts.edb.hkedcity.net/filemanager/file/AandL2chapter/A&L2_(1)%20Lorna.pdf)
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geofrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. ResearchGate. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of*

- Educational Research*, 87(6), 1082–1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: Challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar [Equalization of the dichotomous and polytomus mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: what, why, and how? In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. http://seminar.uny.ac.id/icriems/sites/seminar.uny.ac.id/icriems/files/proceeding2018/ME69_Asma%20Khiyarunnisa.pdf
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <http://etheses.uin-malang.ac.id/14487/1/15110087.pdf>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148.

<https://usnsj.com/index.php/JME/article/view/1607>

- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://www.winsteps.com/winman/copyright.htm>.
- Malhotra, N. K. (2006). *Riset pemasaran*[Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://jurnal.uns.ac.id/ijscs/article/view/49457>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah*[Standards of content for primary and secondary education]. Indonesian Government publicationservice. [https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud Tahun2016 Nomor021.pdf](https://bsnp-indonesia.org/wp-content/uploads/2009/06/Permendikbud_Tahun2016_Nomor021.pdf)
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <http://docplayer.net/21737965-Educational-design-research.html>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://journal.unnes.ac.id/sju/index.php/jere/article/view/46133>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://journal.unnes.ac.id/sju/index.php/ujme/article/view/12643>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya*[Item response theory and its application]. NuhaMedika. <http://staffnew.uny.ac.id/upload/132255129/pendidikan/teori-respons-butir-dan-penerapannya-135hal.pdf>
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran*[Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163.

<http://dx.doi.org/10.30587/postulat.v1i2.2094>

- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst: Journal of Physics Education and Its Application/Karst: Jurnal Pendidikan Fisika dan Terapannya*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://www.jurnal.staidarululumkandangan.ac.id/index.php/annahdhah/article%20/view/40>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: teori tes klasik and respon [Characteristics of items: classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <https://doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan Winstep* [Analysis of research data with rash and Winstep models]. Halaman Moeka Publishing. <http://book.iaincurup.ac.id/index.php/lp2/catalog/download/42/28/122->

[1?inline=1](#)

- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <http://jurnal.untad.ac.id/jurnal/index.php/JEPMT/article/view/14181>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian*[Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student:
Class/Department :
School :

Instructions : Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

- | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1. Given an arithmetic sequence: 3, 8, 13, 18,
The formula for the nth term of the sequence is
A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
C. $U_n = 4n - 1$

Reason: | 2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.
The number of terms in the sequence is...
A. 12 D. 23
B. 13 E. 24
C. 22

Reason: |
| 3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
A. 5 D. 8
B. 6 E. 11
C. 7

Reason: | 4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
A. 308 D. 344
B. 318 E. 354
C. 326

Reason: |
| 5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
Then the middle term of the sequence is ...
A. 53 D. 11
B. 52 E. 10
C. 20
D. 11
E. 10

Reason: | 6. Given the arithmetic sequence: 4, 10, 16, 22,
If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
A. 18 D. 24
B. 20 E. 26
C. 22

Reason: |
| 7. The nth term of an arithmetic series is $U_n = 3n - 5$.
The formula for the sum of the first n terms of the series is ...
A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
C. $S_n = \frac{n}{2}(3n - 4)$

Reason: | 8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....
A. $5n - 20$ D. $2n - 20$
B. $5n - 10$ E. $2n - 10$
C. $2n - 30$

Reason: |
| 9. The sum of all integers between 100 and 300 which are divisible by 5 is ... $S_n = \frac{n}{2}(3n - 7)$
A. 8.200 D. 7.600
B. 8.000 E. 7.400
C. 7.800

Reason: | 10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
A. 24 D. 27
B. 25 E. 28
C. 26

Reason: |
| 11. The middle term of an arithmetic sequence is 25. If the difference is 4 and the 5th term is 21. | 12. A number of candies are distributed among five children according to the rules of an arithmetic |

Then the sum of all the terms in the sequence is ...

- A. 175D.295
- B. 189E.375
- C. 275

Reason:

sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces

- A. 60D.75
- B. 65E.80
- C. 70

Reason:

13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...

- A. 564D.45
- B. 276E.36
- C. 48

Reason:

14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...

- A. 9D.12
- B. 10E.13
- C. 11

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...

- A. 32D.256
- B. 64E.512
- C. 128

Reason:

16. The value of the middle term of the geometric sequence: 6, 3, ..., $3/512$ is ...

- A. $\frac{1}{16}$ D. $\frac{4}{16}$
- B. $\frac{2}{16}$ E. $\frac{5}{16}$
- C. $\frac{3}{16}$

Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm

- A. 18D.35
- B. 24E.40,5
- C. 27,5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...

- A. $\frac{3}{4}$ D. $-\frac{1}{2}$
- B. $-\frac{1}{4}$ E. $-\frac{3}{4}$
- C. $\frac{1}{3}$

Reason:

19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is...m

- A. 60D.90
- B. 70E.100
- C. 80

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.

The value of $3a + b$ is ...

- A. 8D.14
- B. 10E.20
- C. 12

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$.

If $K = L$, then cis...

- A. 12D.15
- B. 13E.16
- C. 14

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$.

Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$
- C. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah ...

- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah ...

- A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...

- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix

- determinant A is ...
 A. 0 D. 2
 B. 1 E. 4
 C. 2

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$.

Inverse matrix AB adalah $(AB)^{-1} = \dots$

- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$

Reason:

33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...

- A. -2 dan $\frac{5}{6}$ D. -2 dan $-\frac{6}{5}$
 B. 2 dan $-\frac{5}{6}$ E. -2 dan $\frac{6}{5}$
 C. 2 dan $\frac{6}{5}$

Reason:

35. The roots of the quadratic equation $2x^2 - 3x - 9 = 0$ are x_1 dan x_2 . Value of $x_1^2 + x_2^2$ is ...

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.

If $A + B = C$, then $x + y = \dots$

- A. -5 D. 3
 B. -1 E. 5
 C. 1

Reason:

26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$

then $A(B - C) = \dots$

- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix.

Value of x that satisfies is ...

- A. -5 D. 3
 B. -4 E. 4
 C. -3

Reason:

30. Transpose matrix P adalah P^t . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^t)^{-1}$ is ...

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$

Reason:

32. The roots of the quadratic equation $3x^2 - 4x - 4$ are ...

- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$

Reason:

34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...

- A. -4 D. 2
 B. -2 E. 4
 C. 0

Reason:

36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α dan β . The quadratic

A. $11\frac{1}{4}$ D. $-6\frac{3}{4}$

B. $6\frac{3}{4}$ E. $-11\frac{1}{4}$

C. $2\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...

A. $y = x^2 - 2x + 1$

B. $y = x^2 - 2x + 3$

C. $y = x^2 + 2x - 1$

D. $y = x^2 + 2x + 1$

E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0, -4)$ then the value of $f(7)$ is ...

A. -16 D. -19

B. -17 E. -20

C. -18

Reason:

equation $(\alpha - 2)$ dan $(\beta - 2)$ is...

a. $x^2 + 6x + 5 = 0$

b. $x^2 + 6x + 7 = 0$

c. $x^2 + 6x + 11 = 0$

d. $x^2 - 2x + 3 = 0$

e. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation ? ...

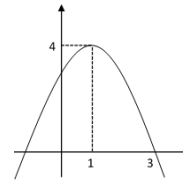
A. $y = x^2 + 2x + 3$

B. $y = x^2 - 2x - 3$

C. $y = -x^2 + 2x - 3$

D. $y = -x^2 - 2x + 3$

E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is...

A. $y = -x^2 + 2x - 3$ D. $y = -x^2 - 2x - 5$

B. $y = -x^2 + 2x + 3$ E. $y = -x^2 - 2x + 5$

C. $y = -x^2 - 2x + 3$

Reason:



SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>

Galley Proof (ID#21112502244011)

2 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>
Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Tue, May 17, 2022 at 6:12 PM

Dear Dr. Sutiarso,

Please see the attached galley proof of your paper (ID#2103030737) (word file). Please highlight in green for your edited parts.

By the way,

- 1- Please check the language of your paper as a proofreading lastly.
- 2- Please check all references regarding with attached citation guide for APA 7 style. (Please see the citation guide page in our web site: <https://www.eujem.com/citation-guide>)

We ask you to check it please. Please edit at word file and resend it to me please in 2 days.

We are looking forward to getting your final paper by **May 19, 2022**.

Best regards,
Ahmet Savas Ph.D.
Editor, European Journal of Educational Management

<http://www.eujem.com>

editor@eujem.com

On 5/10/2022 1:24 PM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

Here I attach a revision of your proofreading suggestion.

Best regards,
Sugeng Sutiarso
Lampung University

On Fri, May 6, 2022 at 7:39 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarso,

Thank you for your kind email. Please see the attached file as the proofreading of your paper.

We are preparing the galley proof of your paper.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 4/23/2022 4:36 AM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

Here, I attach the copyright transfer agreement.

Best regards,
Sugeng Sutiarmo
Lampung University

On Fri, Apr 22, 2022 at 5:24 PM European Journal of Educational Research
<editor@eu-jer.com> wrote:

Dear Dr. Sugeng Sutiarmo,

We have received your payment about your paper entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" ID#21112502244011. Thanks.

We kindly ask from you to sign the copyright transfer agreement for your paper. After all author(s) signed, please scan and send via email to me **as soon as possible**. Please download the pdf file of this agreement from this link : <https://eu-jer.com/EU-JER-copyright-transfer-agreement.pdf> You can use e-signature, if you have. Also you can use your mobil phone as a scanner. If the other author live in another city, he/she sign the paper and send this paper via email. Than you can sign on this paper.

We are preparing the galley proof of your paper. We will send it to you in order to check before publication. The preparing of galley proofs may take some time because of our intensity. Thank you for your patience.

We are looking forward to getting copyright transfer agreement.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com



EU-JER_11_3_1441_SUTIARSO_PROOF.docx

1818K

SUGENG SUTIARSO <sugeng.sutiarmo@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Thu, May 19, 2022 at 9:14 PM

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

I checked the language and references of my paper, and there are some edited parts of the appendix (highlighted in green), namely: (1) the word "and" (previously, the word "dan" in Indonesian) and (2) the word "is" (previously, the word "adalah" in Indonesian).

Here is my final paper attached.

Best regards,
Sugeng Sutiarmo
Lampung University

[Quoted text hidden]



Ok_EU-JER_11_3_1441_SUTIARSO_PROOF.docx

1826K



European Journal of Educational Research

Volume 11, Issue 3, 1441 - 1462.

ISSN: 2165-8714

<https://www.eu-jer.com/>

Developing Assessment Instrument Using Polytomous Response in Mathematics

SugengSutiarso*

University of Lampung, INDONESIA

UndangRosidin

University of Lampung, INDONESIA

AanSulistiawan

Vocational School, INDONESIA

Received: November 25, 2021 • Revised: February 2, 2022 • Accepted: May 17, 2022

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used an open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) the open polytomous response test has a good category according to classical and modern theory. However, the discrimination power of test items in classical theory needs several revisions, (2) the assessment instrument using the polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses.

To cite this article: Sutiarso, S., Rosidin, U., & Sulistiawan, A. (2022). Developing assessment instrument using polytomous response in mathematics. *European Journal of Educational Research*, 11(3), 1441-1462. <https://doi.org/10.12973/eu-jer.11.3.1441>

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools was divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).



Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with specific rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over essay test are firstly, the test can be

* Corresponding author:

SugengSutiarso, University of Lampung. Sumantri Brojonegoro, Bandar Lampung, Indonesia. ✉ sugeng.sutiarso@fkip.unila.ac.id

© 2022 The Author(s). **Open Access**- This article is under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>)



conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weaknesses. The multiple-choice test is not able to portray the actual abilities of students, the answers of this test tend to be guessing games or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). A multiple-choice test with only two answer choices is called a "dichotomous test," and a multiple-choice test with more than two answer choices is called a "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or henceforth called as polytomous response test (Suwarto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as a test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), a test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into an open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response tests for students in college and senior or junior school. Students in college and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools are not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomous test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmalwati, 2020). Therefore, the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomous response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing the item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because, in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows: (1) Does

the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory? and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's (2013) model, with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

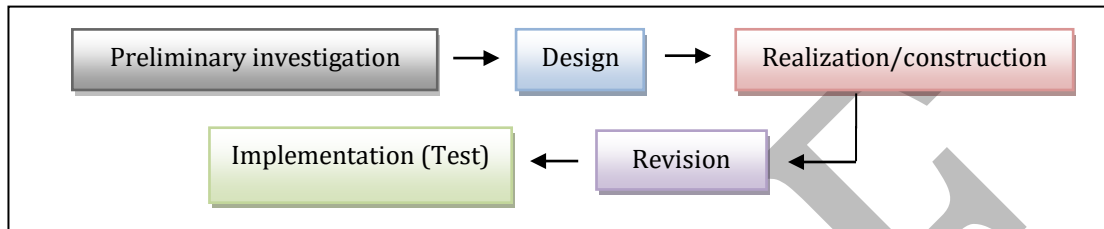


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the

indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. Questionnaire data analysis (qualitative analysis)

There are two sets of questionnaire data, namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. Test data analysis (empirical analysis)

After conducting the content validity test, the researcher conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to have good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's Alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program was used for classical theory

and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep) because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Correlation column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit's MNSQ value is 0.5 to 1.5 and the Pt-Measure Correlation is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay tests and 20% of teachers used multiple-choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

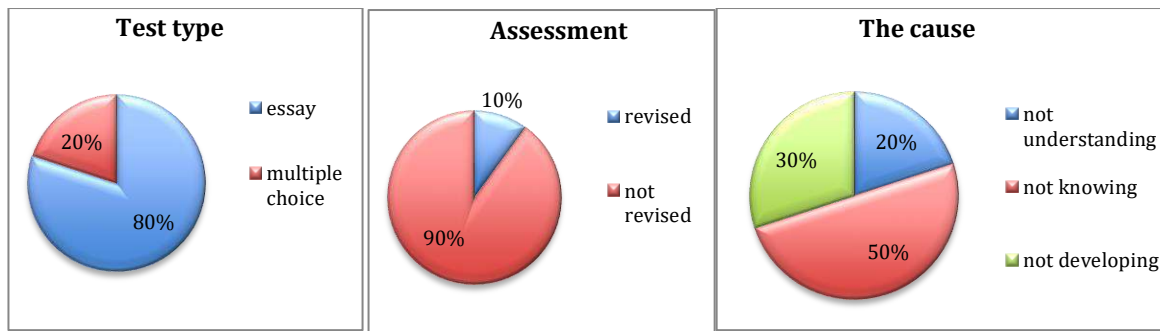


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	Good	0.196	Revised	21	0.482	Good	0.143	Revised
2	0.487	Good	0.179	Revised	22	0.535	Good	0.429	Good
3	0.528	Good	0.214	Revised	23	0.492	Good	0.250	Revised
4	0.540	Good	0.304	Good	24	0.438	Good	-0.071	Revised
5	0.489	Good	0.089	Revised	25	0.436	Good	-0.107	Revised
6	0.446	Good	-0.161	Revised	26	0.385	Good	-0.286	Revised
7	0.438	Good	-0.232	Revised	27	0.383	Good	-0.321	Revised
8	0.453	Good	-0.143	Revised	28	0.416	Good	-0.143	Revised
9	0.414	Good	-0.143	Revised	29	0.458	Good	-0.125	Revised
10	0.409	Good	-0.339	Revised	30	0.385	Good	-0.375	Revised
11	0.438	Good	-0.143	Revised	31	0.404	Good	-0.321	Revised
12	0.436	Good	-0.036	Revised	32	0.433	Good	-0.250	Revised
13	0.400	Good	-0.321	Revised	33	0.441	Good	0,036	Revised
14	0.450	Good	-0.036	Revised	34	0.424	Good	-0.268	Revised
15	0.462	Good	0.250	Revised	35	0.412	Good	-0.321	Revised
16	0.453	Good	-0.089	Revised	36	0.431	Good	-0.304	Revised
17	0.416	Good	-0.143	Revised	37	0.404	Good	-0.232	Revised
18	0.419	Good	-0.196	Revised	38	0.363	Good	-0.482	Revised
19	0.431	Good	-0.232	Revised	39	0.230	Good	-0.929	Revised
20	0.441	Good	-0.089	Revised	40	0.211	Good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

*Analysis of Test Data with Modern Theory**The Unidimensional Assumption Test*

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first-factor Eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first-factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

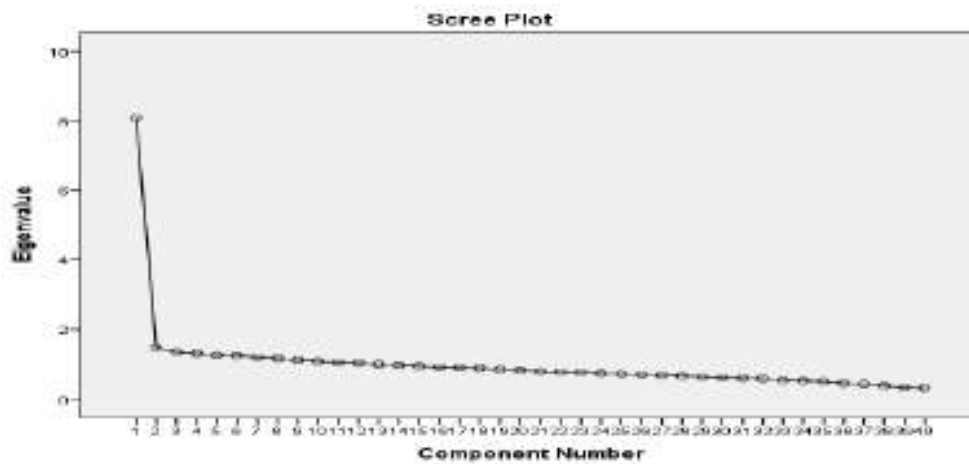


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Correlation is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	-.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	-.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	-.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	-.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained were presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	-.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	-.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	-.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	-.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

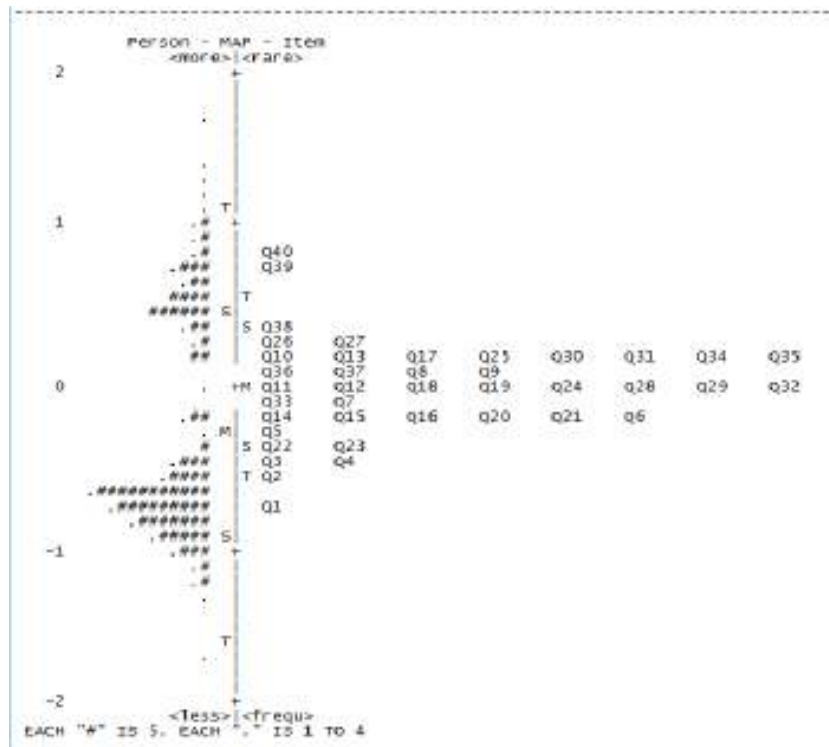


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (PT-Measure Correlation column). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

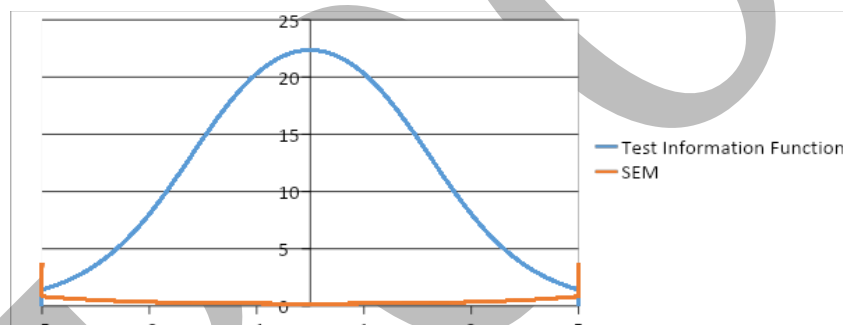


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in the vocational school was based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

Question 1:

Given an arithmetic sequence: 3, 8, 13, 18,
The formula for the nth term of the sequence is ...

A. $U_n = 5n - 3$
B. $U_n = 5n - 2$
C. $U_n = 2n + 1$
D. $U_n = 4n - 1$
E. $U_n = 3n + 2$

Reason:

Pattern 1:

Pattern 2:

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term) and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

Question 2:

Given an arithmetic sequence: 2, 5, 8, 11,, 68.
The number of terms in the sequence is...

A. 12
B. 13
C. 22
D. 23
E. 24

Reason:

Pattern 1:

Pattern 2:

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

Question 3:

An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...

A. 5
B. 6
C. 7
D. 8
E. 11

Reason:

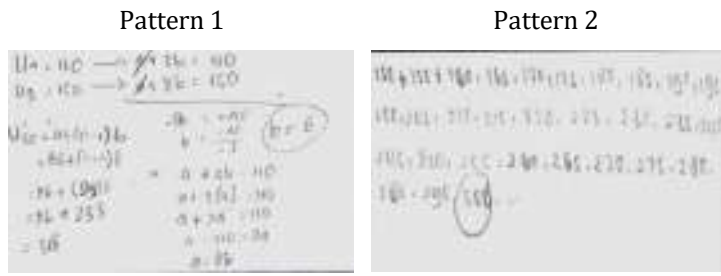
Pattern 1:

Pattern 2:

Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

Question 4:
 The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308
 B. 318
 C. 326
 D. 344
 E. 354

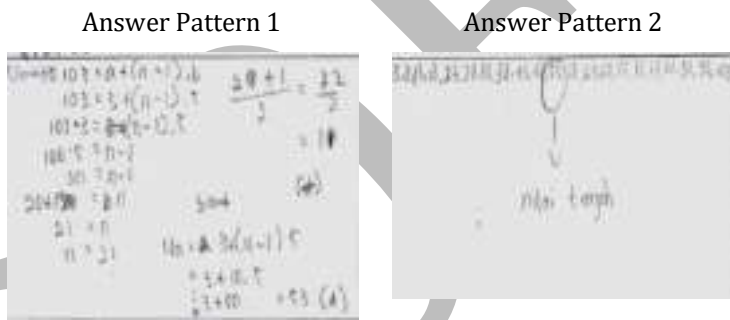


Reason:

Figure 13. An Example of Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

Question 5:
 An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
 Then the middle term of the sequence is ...308
 A. 53
 B. 52
 C. 20
 D. 11
 E. 10



Reason:

Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice test is easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score. Then, the result is multiplied by 10 to get a value range of 0–10 or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomus response test is more secondhand and complete than other assessment instruments, such as multiple-choice tests(Gierl et al., 2017) or essay tests(Putri et al., 2020).

Discussion

This research is development research aimed at developing a good mathematical assessment instrument using polytomous responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item

discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of the classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to the modern theory before revising or replacing the item.

Research on learning assessment with open response polytomous was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

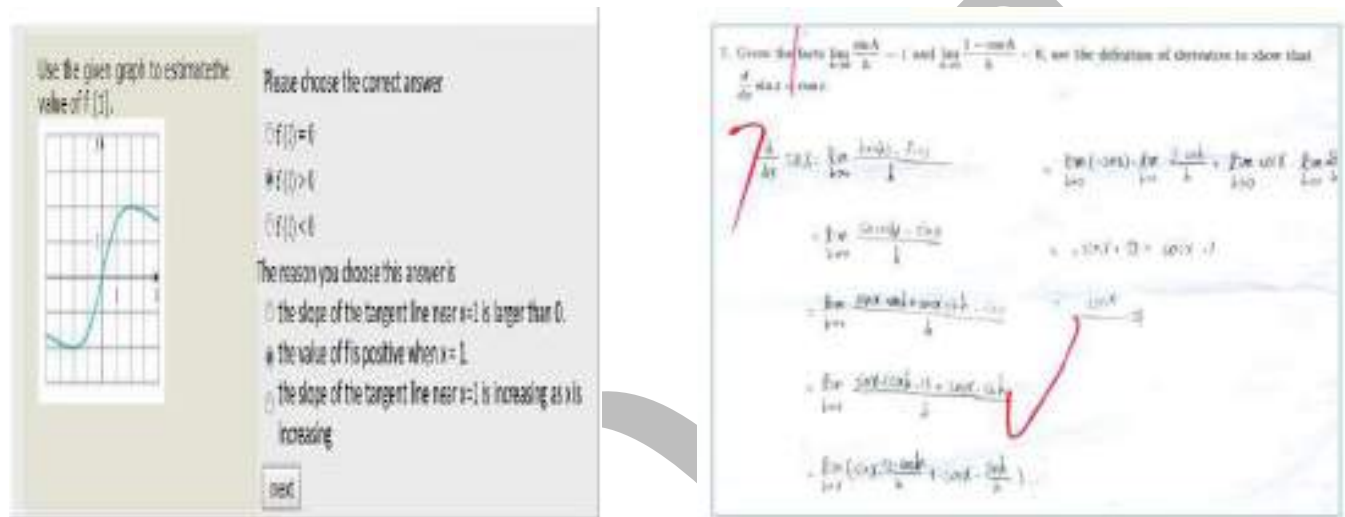


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous

Response Test

Another study on learning assessment with open response polychromous was conducted by Ayanwale(2021).Ayanwale's research compares two polytomous response test analysis methods namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

Other studies related to classical and modern theory were conducted bySarea (2018)and Saepuzaman et al.(2021). Sarea's research states that the response polytomous test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomous test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulasand (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

- A. 308
- B. 318
- C. 326
- D. 344
- E. 354

Reason:

$$\begin{aligned}
 U_4 &= 110 \rightarrow a + 3b = 110 \\
 U_9 &= 150 \rightarrow a + 8b = 150 \\
 \hline
 -5b &= -40 \\
 b &= \frac{-40}{-5} \quad (b = 8) \\
 a + 3(8) &= 110 \\
 a + 24 &= 110 \\
 a &= 110 - 24 \\
 a &= 86 \\
 U_{30} &= a + (n-1)b \\
 &= 86 + (30-1)8 \\
 &= 86 + 233 \\
 &= 318
 \end{aligned}$$

(i)

(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, using a drawing or model, analogy, and formula. Syahlan(2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syahlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. Thus, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further

research, it is suggested to conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

Authorship Contribution Statement

Sutiarso: Conceptualization, design, development of instruments, analysis, article writing, final approval.

Rosidin: Design, development of instruments, analysis, editing. Sulistiawan: Development of instruments, collect data, analysis, editing.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://bit.ly/39wv73P>
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomis dan politomis generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *JurnalSuluh Pendidikan*, 17(1), 32-37. <https://bit.ly/3yFQ6eO>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://bit.ly/3Nob4mN>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://bit.ly/3lfARkW>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. <https://bit.ly/3MnA9hw>
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geoffrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. ResearchGate. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. <https://doi.org/10.3102/0034654317726529>

- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: Challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523-545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar [Equalization of the dichotomous and polytomus mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: What, why, and how? In A. W. Subianto (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <https://bit.ly/3liFMS8>
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <https://bit.ly/3sEcNMA>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://bit.ly/3Psrc8s>
- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://bit.ly/3wljq95>
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://bit.ly/3yIePzb>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Standards of content for primary and secondary education]. Indonesian Government publicationservice. <https://bit.ly/3FZHlhp>
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <https://bit.ly/3MzqkNm>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://bit.ly/37Qctyu>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://bit.ly/3yF4iF5>

- Retnawati, H. (2014). *Teori respons butir dan penerapannya* [Item response theory and its application]. NuhaMedika. <https://bit.ly/39TFDIF>
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163. <https://doi.org/10.30587/postulat.v1i2.2094>
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst: Journal of Physics Education and Its Application/ Karst: Jurnal Pendidikan Fisika dan Terapannya*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://bit.ly/3FTOU7>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Teori tes klasik and respon [Characteristics of items: Classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunika.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <https://doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan Winstep* [Analysis of research data with rash and Winstep models]. Halaman Moeka Publishing. <https://bit.ly/3NlnGLE>
- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <https://bit.ly/3wjuXp9>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/ Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student:

Class/Department :

School :

Instructions: Mark (x) one of the correct answer choices, and give the reason

(use another piece of paper to write down your reason)

1. Given an arithmetic sequence: 3, 8, 13, 18, ...
The formula for the n th term of the sequence is ...
A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
C. $U_n = 4n - 1$
Reason:
2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.
The number of terms in the sequence is...
A. 12 D. 23
B. 13 E. 24
C. 22
Reason:
3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
A. 5 D. 8
B. 6 E. 11
C. 7
Reason:
4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
A. 308 D. 344
B. 318 E. 354
C. 326
Reason:
5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
Then the middle term of the sequence is ...
A. 53 D. 11
B. 52 E. 10
C. 20
D. 11
E. 10
Reason:
6. Given the arithmetic sequence: 4, 10, 16, 22, ...
If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
A. 18 D. 24
B. 20 E. 26
C. 22
Reason:
7. The n th term of an arithmetic series is $U_n = 3n - 5$.
The formula for the sum of the first n terms of the series is ...
A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
C. $S_n = \frac{n}{2}(3n - 4)$
Reason:
8. The sum of the first n terms of an arithmetic series.
 $S_n = n^2 - 19n$. The formula for the n th term of the Series is.....
A. $5n - 20$ D. $2n - 20$
B. $5n - 10$ E. $2n - 10$
C. $2n - 30$
Reason:
9. The sum of all integers between 100 and 300 which are divisible by 5 is ... $S_n = \frac{n}{2}(3n - 7)$
A. 8.200 D. 7.600
10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?

- B. 8.000 E. 7.400
- C. 7.800

Reason:

- A. 24 D. 27
- B. 25 E. 28
- C. 26

Reason:

11. The middle term of an arithmetic sequence is 25.
If the difference is 4 and the 5th term is 21.
Then the sum of all the terms in the sequence is ...
- A. 175 D. 295
 - B. 189 E. 375
 - C. 275

Reason:

12. A number of candies are distributed among five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces
- A. 60 D. 75
 - B. 65 E. 80
 - C. 70

Reason:

13. The sum of the first n terms of a series is $2n^2-n$.
So the 12th term of the series is...
- A. 564 D. 45
 - B. 276 E. 36
 - C. 48

Reason:

14. The number of terms in the geometric sequence:
3, 6, 12, ..., 3072 is ...
- A. 9 D. 12
 - B. 10 E. 13
 - C. 11

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...
- A. 32 D. 256
 - B. 64 E. 512
 - C. 128

Reason:

16. The value of the middle term of the geometric sequence:
6, 3, ..., $3/512$ is ...

- A. $\frac{1}{16}$ D. $\frac{4}{16}$
- B. $\frac{2}{16}$ E. $\frac{5}{16}$
- C. $\frac{3}{16}$

Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm
- A. 18 D. 35
 - B. 24 E. 40,5
 - C. 27,5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...
- A. $\frac{3}{4}$ D. $-\frac{1}{2}$
 - B. $\frac{1}{4}$ E. $-\frac{3}{4}$
 - C. $\frac{1}{3}$

Reason:

19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is.... m
- A. 60 D. 90
 - B. 70 E. 100
 - C. 80

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.
The value of $3a + b$ is ...
- A. 8 D. 14
 - B. 10 E. 20
 - C. 12

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$.

If $K = L$, then cis ...

- A. 12 D. 15
- B. 13 E. 16
- C. 14

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ adalah ...

- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
- B. $\begin{bmatrix} 1 & 5 & 6 \\ 3 & 3 & 7 \\ 1 & 3 & 1 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
- C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ adalah ...

- A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
- B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$
- C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...

- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
- B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
- C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ...

- A. 0 D. 2
- B. 1 E. 4
- C. 2

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}, C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$.

Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$
- C. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$

Reason:

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}, B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.

If $A + B = C$, then $x + y = ...$

- A. -5 D. 3
- B. -1 E. 5
- C. 1

Reason:

26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}, B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$

then $A(B - C) = ...$

- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
- B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
- C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix. Value of x that satisfies is ...

- A. -5 D. 3
- B. -4 E. 4
- C. -3

Reason:

30. Transpose matrix P adalah P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ...

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
- B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
- C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$.
 Inverse matrix AB adalah $(AB)^{-1} = \dots$
 A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$
 Reason:

32. The roots of the quadratic equation $3x^2 - 4x - 4 = 0$ are ...
 A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$
 Reason:

33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...
 A. -2 dan $\frac{5}{6}$ D. -2 dan $-\frac{6}{5}$
 B. 2 dan $-\frac{5}{6}$ E. -2 dan $\frac{6}{5}$
 C. 2 dan $\frac{6}{5}$
 Reason:

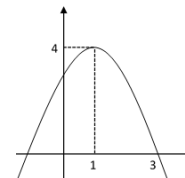
34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...
 A. -4 D. 2
 B. -2 E. 4
 C. 0
 Reason:

35. The roots of the quadratic equation $2x^2 - 3x - 9 = 0$ are x_1 dan x_2 . Value of $x_1^2 + x_2^2$ is ...
 A. $11\frac{1}{4}$ D. $-6\frac{3}{4}$
 B. $6\frac{3}{4}$ E. $-11\frac{1}{4}$
 C. $2\frac{1}{4}$
 Reason:

36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α dan β . The quadratic equation $(\alpha - 2)$ dan $(\beta - 2)$ is ...
 a. $x^2 + 6x + 5 = 0$
 b. $x^2 + 6x + 7 = 0$
 c. $x^2 + 6x + 11 = 0$
 d. $x^2 - 2x + 3 = 0$
 e. $x^2 + 2x + 11 = 0$
 Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...
 A. $y = x^2 - 2x + 1$
 B. $y = x^2 - 2x + 3$
 C. $y = x^2 + 2x - 1$
 D. $y = x^2 + 2x + 1$
 E. $y = x^2 + 2x + 3$
 Reason:

38. The figure below is a graph of the quadratic equation? ...
 A. $y = x^2 + 2x + 3$
 B. $y = x^2 - 2x - 3$
 C. $y = -x^2 + 2x - 3$
 D. $y = -x^2 - 2x + 3$
 E. $y = -x^2 + 2x + 3$
 Reason:



39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0, -4)$ then the value of $f(7)$ is ...
 A. -16 D. -19
 B. -17 E. -20
 C. -18
 Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...
 A. $y = -x^2 + 2x - 3$ D. $y = -x^2 - 2x - 5$
 B. $y = -x^2 + 2x + 3$ E. $y = -x^2 - 2x + 5$
 C. $y = -x^2 - 2x + 3$
 Reason:



European Journal of Educational Research

Volume 11, Issue 3, 1441 - 1462.

ISSN: 2165-8714

<https://www.eu-jer.com/>

Developing Assessment Instrument Using Polytomous Response in Mathematics

Sugeng Sutiarto* 

University of Lampung, INDONESIA

Undang Rosidin 

University of Lampung, INDONESIA

Aan Sulistiawan

Vocational School, INDONESIA

Received: November 25, 2021 • Revised: February 2, 2022 • Accepted: May 17, 2022

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used an open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) the open polytomous response test has a good category according to classical and modern theory. However, the discrimination power of test items in classical theory needs several revisions, (2) the assessment instrument using the polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: *Assessment instrument, classical and modern theory, vocational school, polytomous responses.*

To cite this article: Sutiarto, S., Rosidin, U., & Sulistiawan, A. (2022). Developing assessment instrument using polytomous response in mathematics. *European Journal of Educational Research*, 11(3), 1441-1462. <https://doi.org/10.12973/eu-jer.11.3.1441>

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools was divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).

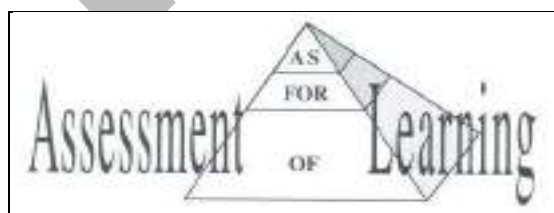


Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with specific rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over essay test are firstly, the test can be

*Corresponding author:

Sugeng Sutiarto, University of Lampung. Sumantri Brojonegoro, Bandar Lampung, Indonesia. ✉ sugeng.sutiarto@fkip.unila.ac.id



conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weaknesses. The multiple-choice test is not able to portray the actual abilities of students, the answers of this test tend to be guessing games or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). A multiple-choice test with only two answer choices is called a "dichotomous test," and a multiple-choice test with more than two answer choices is called a "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or hence forth called as polytomous response test (Suwanto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as a test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), a test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into an open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response tests for students in college and senior or junior school. Students in college and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools is not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomus test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmawati, 2020). Therefore, the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomus response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing the item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because, in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows: (1) Does

the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory? and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's (2013) model, with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

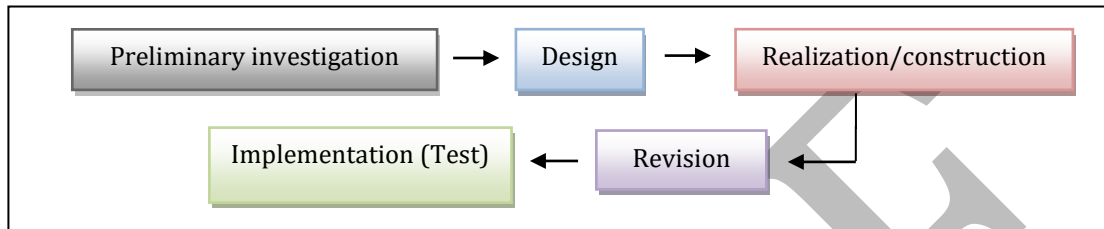


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the

indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. Questionnaire data analysis (qualitative analysis)

There are two sets of questionnaire data namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. Test data analysis (empirical analysis)

After conducting the content validity test, the researchers conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to having good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program was used for classical theory

and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep) because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Correlation column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit's MNSQ value is 0.5 to 1.5 and the Pt-Measure Correlation is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Findings/Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay tests and 20% of teachers used multiple-choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

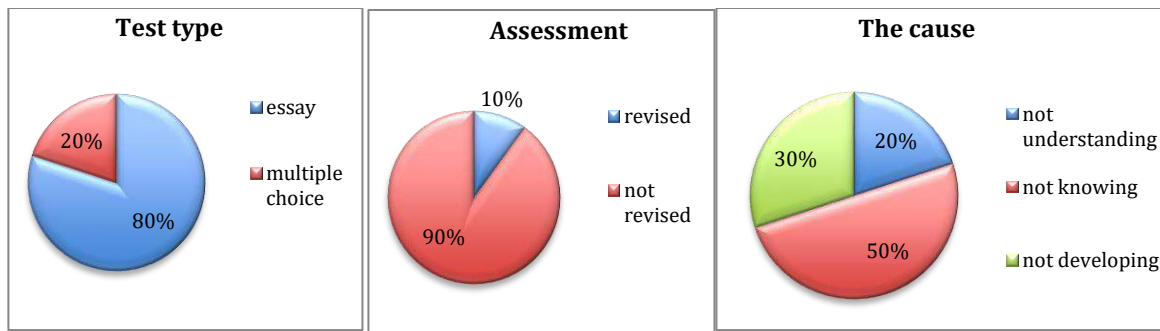


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Bartlett's Test of Sphericity	Approx. Chi-Square
	Df
	Sig.
	1936.378
	780
	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	Good	0.196	Revised	21	0.482	Good	0.143	Revised
2	0.487	Good	0.179	Revised	22	0.535	Good	0.429	Good
3	0.528	Good	0.214	Revised	23	0.492	Good	0.250	Revised
4	0.540	Good	0.304	Good	24	0.438	Good	-0.071	Revised
5	0.489	Good	0.089	Revised	25	0.436	Good	-0.107	Revised
6	0.446	Good	-0.161	Revised	26	0.385	Good	-0.286	Revised
7	0.438	Good	-0.232	Revised	27	0.383	Good	-0.321	Revised
8	0.453	Good	-0.143	Revised	28	0.416	Good	-0.143	Revised
9	0.414	Good	-0.143	Revised	29	0.458	Good	-0.125	Revised
10	0.409	Good	-0.339	Revised	30	0.385	Good	-0.375	Revised
11	0.438	Good	-0.143	Revised	31	0.404	Good	-0.321	Revised
12	0.436	Good	-0.036	Revised	32	0.433	Good	-0.250	Revised
13	0.400	Good	-0.321	Revised	33	0.441	Good	0,036	Revised
14	0.450	Good	-0.036	Revised	34	0.424	Good	-0.268	Revised
15	0.462	Good	0.250	Revised	35	0.412	Good	-0.321	Revised
16	0.453	Good	-0.089	Revised	36	0.431	Good	-0.304	Revised
17	0.416	Good	-0.143	Revised	37	0.404	Good	-0.232	Revised
18	0.419	Good	-0.196	Revised	38	0.363	Good	-0.482	Revised
19	0.431	Good	-0.232	Revised	39	0.230	Good	-0.929	Revised
20	0.441	Good	-0.089	Revised	40	0.211	Good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

*Analysis of Test Data with Modern Theory**The Unidimensional Assumption Test*

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first-factor Eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first-factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

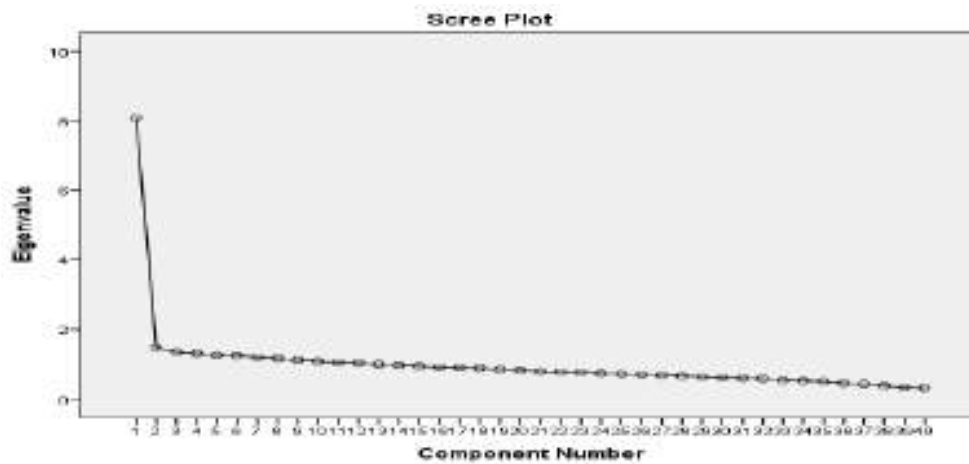


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Correlation is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ in fit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	-.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	-.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained were presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	-.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	-.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

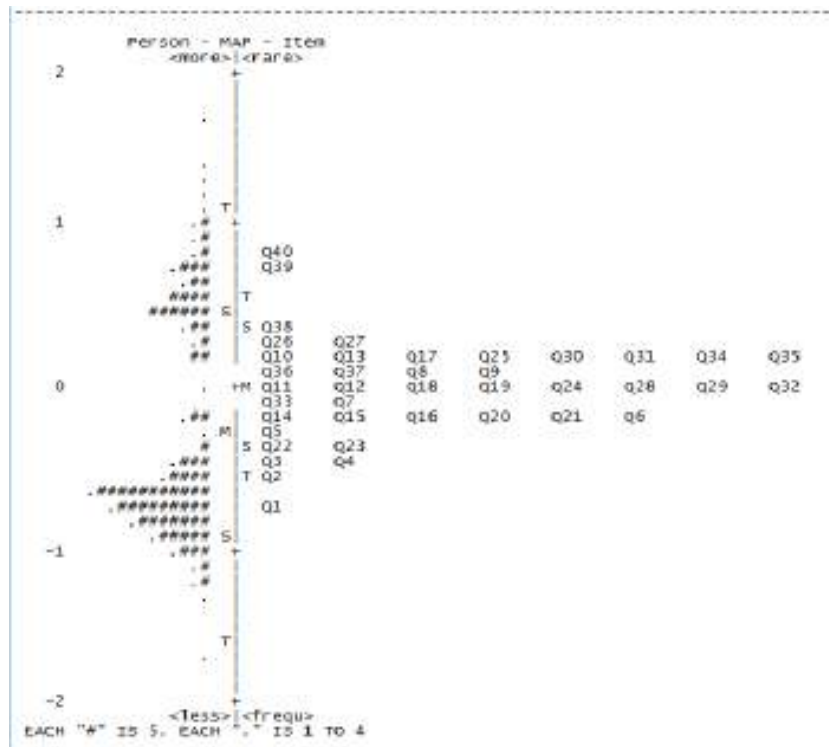


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (PT-Measure Correlation column). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

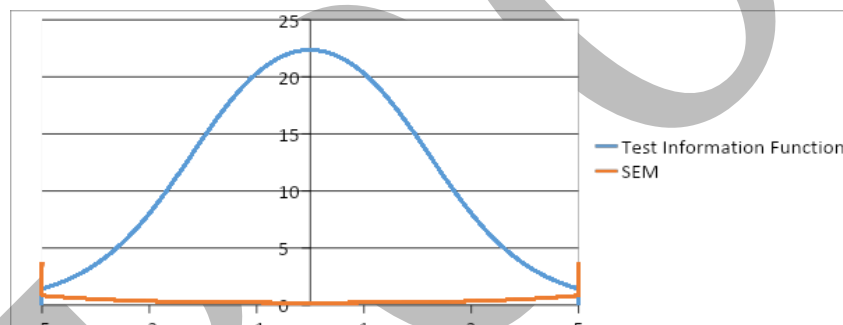


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in the vocational school was based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

Question 1:

Given an arithmetic sequence: 3, 8, 13, 18,
The formula for the nth term of the sequence is ...

A. $U_n = 5n - 3$
B. $U_n = 5n - 2$
C. $U_n = 2n + 1$
D. $U_n = 4n - 1$
E. $U_n = 3n + 2$

Reason:

Pattern 1:

Pattern 2:

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term) and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

Question 2:

Given an arithmetic sequence: 2, 5, 8, 11, ..., 68.
The number of terms in the sequence is ...

A. 12
B. 13
C. 22
D. 23
E. 24

Reason:

Pattern 1:

Pattern 2:

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

Question 3:

An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is ...

A. 5
B. 6
C. 7
D. 8
E. 11

Reason:

Pattern 1:

Pattern 2:

Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

Question 4:
 The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308
 B. 318
 C. 326
 D. 344
 E. 354

Reason:

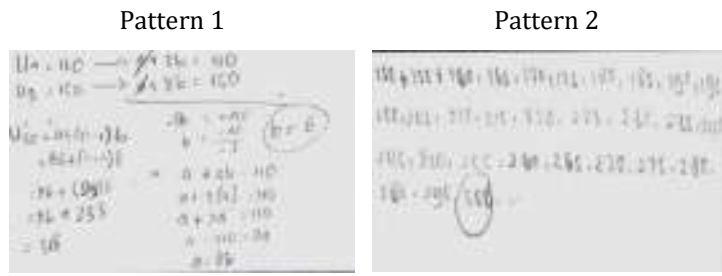


Figure 13. An Example of Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

Question 5:
 An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
 Then the middle term of the sequence is ...308
 A. 53
 B. 52
 C. 20
 D. 11
 E. 10

Reason:

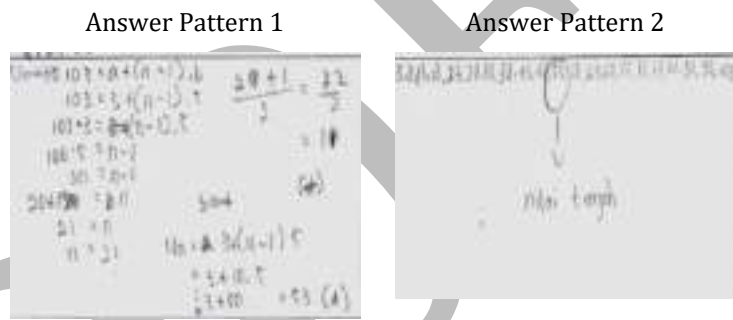


Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice test is easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score. Then, the result is multiplied by 10 to get a value range of 0–10 or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomus response test is more secondhand and complete than other assessment instruments, such as multiple-choice tests (Gierl et al., 2017) or essay tests (Putri et al., 2020).

Discussion

This research is development research aimed at developing a good mathematical assessment instrument using polytomous responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item

discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of the classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to the modern theory before revising or replacing the item.

Research on learning assessment with open response polytomous was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

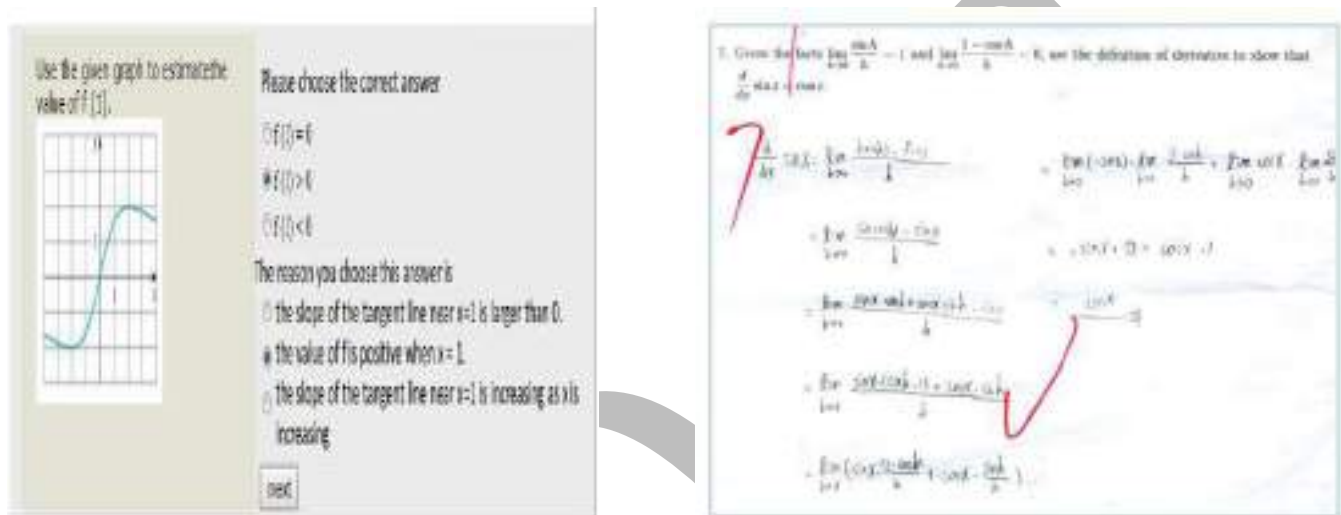


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous

Response Test

Another study on learning assessment with open response polychromous was conducted by Ayanwale (2021). Ayanwale's research compares two polytomous response test analysis methods namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

Other studies related to classical and modern theory were conducted by Sarea (2018) and Saepuzaman et al. (2021). Sarea's research states that the response polytomous test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomous test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

- A. 308
- B. 318
- C. 326
- D. 344
- E. 354

Reason:

$$\begin{aligned}
 U_4 &= 110 \rightarrow a + 3b = 110 \\
 U_9 &= 150 \rightarrow a + 8b = 150 \\
 \hline
 -5b &= -40 \\
 b &= \frac{-40}{-5} \quad (b = 8) \\
 a + 3(8) &= 110 \\
 a + 24 &= 110 \\
 a &= 110 - 24 \\
 a &= 86 \\
 U_{30} &= a + (n-1)b \\
 &= 86 + (30-1)8 \\
 &= 86 + 233 \\
 &= 318
 \end{aligned}$$

(i)

(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, using a drawing or model, analogy, and formula. Syahlan (2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syahlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. Thus, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further

research, it is suggested to conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

Authorship Contribution Statement

Sutiarso: Conceptualization, design, development of instruments, analysis, article writing, final approval. Rosidin: Design, development of instruments, analysis, editing. Sulistiawan: Development of instruments, collect data, analysis, editing.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://bit.ly/39wv73P>
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomis dan politomis generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *Jurnal Suluh Pendidikan*, 17(1), 32-37. <https://bit.ly/3yFQ6eO>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://bit.ly/3Nob4mN>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://bit.ly/3lfARKW>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. <https://bit.ly/3MnA9hw>
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geofrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. Research Gate. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. <https://doi.org/10.3102/0034654317726529>

- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: Challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>
- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523-545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar [Equalization of the dichotomous and polytomus mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: What, why, and how? In A. W. Subianto (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <https://bit.ly/3liFMS8>
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <https://bit.ly/3sEcNMA>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://bit.ly/3Psrc8s>
- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://bit.ly/3wliq95>
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://bit.ly/3yIePzb>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Standards of content for primary and secondary education]. Indonesian Government publication service. <https://bit.ly/3FZHLhp>
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <https://bit.ly/3MzqkNm>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://bit.ly/37QCtyu>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://bit.ly/3yF4iF5>

- Retnawati, H. (2014). *Teori respons butir dan penerapannya* [Item response theory and its application]. Nuha Medika. <https://bit.ly/39TFDIF>
- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163. <https://doi.org/10.30587/postulat.v1i2.2094>
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst: Journal of Physics Education and Its Application/Karst: Jurnal Pendidikan Fisika dan Terapannya*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://bit.ly/3FToOU7>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Teori tes klasik and respon [Characteristics of items: Classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <https://doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan Winstep* [Analysis of research data with rash and Winstep models]. Halaman Moeka Publishing. <https://bit.ly/3NlnGLE>
- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <https://bit.ly/3wjuXp9>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student:

Class/Department :

School :

Instructions: Mark (x) one of the correct answer choices, and give the reason

(use another piece of paper to write down your reason)

1. Given an arithmetic sequence: 3, 8, 13, 18, The formula for the n th term of the sequence is
 A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
 B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
 C. $U_n = 4n - 1$
 Reason:
2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...
 A. 12 D. 23
 B. 13 E. 24
 C. 22
 Reason:
3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
 A. 5 D. 8
 B. 6 E. 11
 C. 7
 Reason:
4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308 D. 344
 B. 318 E. 354
 C. 326
 Reason:
5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...
 A. 53 D. 11
 B. 52 E. 10
 C. 20
 D. 11
 E. 10
 Reason:
6. Given the arithmetic sequence: 4, 10, 16, 22, If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
 A. 18 D. 24
 B. 20 E. 26
 C. 22
 Reason:
7. The n th term of an arithmetic series is $U_n = 3n - 5$. The formula for the sum of the first n terms of the series is ...
 A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
 B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
 C. $S_n = \frac{n}{2}(3n - 4)$
 Reason:
8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the n th term of the Series is.....
 A. $5n - 20$ D. $2n - 20$
 B. $5n - 10$ E. $2n - 10$
 C. $2n - 30$
 Reason:
9. The sum of all integers between 100 and 300 which are divisible by 5 is ...
 A. 8.200 D. 7.600
 B. 8.000 E. 7.400
 C. 7.800
 Reason:
10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
 A. 24 D. 27
 B. 25 E. 28
 C. 26
 Reason:

11. The middle term of an arithmetic sequence is 25. If the difference is 4 and the 5th term is 21. Then the sum of all the terms in the sequence is ...

- A. 175
- B. 189
- C. 275
- D. 295
- E. 375

Reason:

12. A number of candies are distributed among five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces

- A. 60
- B. 65
- C. 70
- D. 75
- E. 80

Reason:

13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...

- A. 564
- B. 276
- C. 48
- D. 45
- E. 36

Reason:

14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...

- A. 9
- B. 10
- C. 11
- D. 12
- E. 13

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...

- A. 32
- B. 64
- C. 128
- D. 256
- E. 512

Reason:

16. The value of the middle term of the geometric sequence: 6, 3, ..., $3/512$ is ...

- A. $\frac{1}{16}$
- B. $\frac{2}{16}$
- C. $\frac{3}{16}$
- D. $\frac{4}{16}$
- E. $\frac{5}{16}$

Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm

- A. 18
- B. 24
- C. 27,5
- D. 35
- E. 40,5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...

- A. $\frac{3}{4}$
- B. $\frac{1}{4}$
- C. $\frac{1}{3}$
- D. $-\frac{1}{2}$
- E. $-\frac{3}{4}$

Reason:

19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is... m

- A. 60
- B. 70
- C. 80
- D. 90
- E. 100

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.

The value of $3a + b$ is ...

- A. 8
- B. 10
- C. 12
- D. 14
- E. 20

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$. If $K = L$, then c is ...

- A. 12
- B. 13
- C. 14
- D. 15
- E. 16

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$. Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$
- D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$

is ...

- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$

- D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 E. $\begin{bmatrix} 1 & 3 & 1 \\ 0 & 7 & 6 \\ 2 & 4 & 5 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ is...

- A. $\begin{bmatrix} 13 & 42 \\ 26 & 84 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \\ 26 & 42 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$

- D. $\begin{bmatrix} 13 & 84 \\ 30 & 36 \end{bmatrix}$
 E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$

Reason:

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$,
 and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.

If $A + B = C$, then $x + y = \dots$

- A. -5
 B. -1
 C. 1
 D. 3
 E. 5

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$

is ...

- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$

- D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix.

Value of x that satisfies is ...

- A. -5
 B. -4
 C. -3
 D. 3
 E. 4

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then

matrix determinant A is ...

- A. 0
 B. 1
 C. 2
 D. 2
 E. 4

Reason:

30. Transpose matrix P is P^t . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then
 matrix $(P^t)^{-1}$ is ...

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$
 D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$.

Inverse matrix $(AB)^{-1} = \dots$

- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$
 D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$

Reason:

32. The roots of the quadratic equation $3x^2 - 4x + 3 = 0$ are ...

- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$

Reason:

33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...
- A. -2 and $\frac{5}{6}$
 - B. 2 and $-\frac{5}{6}$
 - C. 2 and $\frac{6}{5}$
 - D. -2 and $-\frac{6}{5}$
 - E. -2 and $\frac{6}{5}$

Reason:

35. The roots of the quadratic equation: $2x^2 - 3x - 9 = 0$ are x_1 and x_2 . Value of $x_1^2 + x_2^2$ is ...
- A. $11\frac{1}{4}$
 - B. $6\frac{3}{4}$
 - C. $2\frac{1}{4}$
 - D. $-6\frac{3}{4}$
 - E. $-11\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...
- A. $y = x^2 - 2x + 1$
 - B. $y = x^2 - 2x + 3$
 - C. $y = x^2 + 2x - 1$
 - D. $y = x^2 + 2x + 1$
 - E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0,-4)$ then the value of $f(7)$ is ...
- A. -16
 - B. -17
 - C. -18
 - D. -19
 - E. -20

Reason:

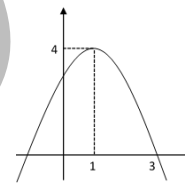
34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...
- A. -4
 - B. -2
 - C. 0
 - D. 2
 - E. 4

Reason:

36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α and β . The quadratic equation $(\alpha - 2)$ and $(\beta - 2)$ is ...
- a. $x^2 + 6x + 5 = 0$
 - b. $x^2 + 6x + 7 = 0$
 - c. $x^2 + 6x + 11 = 0$
 - d. $x^2 - 2x + 3 = 0$
 - e. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation? ...
- A. $y = x^2 + 2x + 3$
 - B. $y = x^2 - 2x - 3$
 - C. $y = -x^2 + 2x - 3$
 - D. $y = -x^2 - 2x + 3$
 - E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...
- A. $y = -x^2 + 2x - 3$
 - B. $y = -x^2 + 2x + 3$
 - C. $y = -x^2 - 2x + 3$
 - D. $y = -x^2 - 2x - 5$
 - E. $y = -x^2 - 2x + 5$

Reason:



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Final Paper (ID#21112502244011)

2 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>
Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Sat, May 21, 2022 at 3:44 AM

Dear Dr. Sutiarso,

Thank you for your email. We have updated your paper. We have corrected a few mistakes.

Please find the attached finalized paper will be published.

We will publish it as online first soon.

The official publication date of your paper is July 15, 2022.

Best regards,
Ahmet Savas Ph.D.
Editor- European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 5/19/2022 5:14 PM, SUGENG SUTJARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

I checked the language and references of my paper, and there are some edited parts of the appendix (highlighted in green), namely: (1) the word "and" (previously, the word "dan" in Indonesian) and (2) the word "is" (previously, the word "adalah" in Indonesian).
Here is my final paper attached.

Best regards,
Sugeng Sutiarso
Lampung University

On Tue, May 17, 2022 at 6:12 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarso,

Please see the attached galley proof of your paper (ID#2103030737) (word file). Please highlight in green for your edited parts.

By the way,

- 1- Please check the language of your paper as a proofreading lastly.
- 2- Please check all references regarding with attached citation guide for APA 7 style. (Please see the citation guide page in our web site: <https://www.eujem.com/citation-guide>)

We ask you to check it please. Please edit at word file and resend it to me please in 2 days.

We are looking forward to getting your final paper by **May 19, 2022**.

Best regards,
Ahmet Savas Ph.D.
Editor, European Journal of Educational Management

<http://www.eujem.com>

editor@eujem.com

On 5/10/2022 1:24 PM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

Here I attach a revision of your proofreading suggestion.

Best regards,
Sugeng Sutiarmo
Lampung University

On Fri, May 6, 2022 at 7:39 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarmo,

Thank you for your kind email. Please see the attached file as the proofreading of your paper.

We are preparing the galley proof of your paper.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 4/23/2022 4:36 AM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

Here, I attach the copyright transfer agreement.

Best regards,
Sugeng Sutiarmo
Lampung University

On Fri, Apr 22, 2022 at 5:24 PM European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sugeng Sutiarmo,

We have received your payment about your paper entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" ID#21112502244011. Thanks.


We kindly ask from you to sign the copyright transfer agreement for your paper. After all author(s) signed, please scan and send via email to me **as soon as possible**. Please download the pdf file of this agreement from this link : <https://eu-jer.com/EU-JER-copyright-transfer-agreement.pdf> You can use e-signature, if you have. Also you can use your mobil phone as a scanner. If the other author live in another city, he/she sign the paper and send this paper via email. Than you can sign on this paper.

We are preparing the galley proof of your paper. We will send it to you in order to check before publication. The preparing of galley proofs may take some time because of our intensity. Thank you for your patience.

We are looking forward to getting copyright transfer agreement.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

 EU-JER_11_3_1441_SUTIARSO_FINAL.docx
1804K

SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

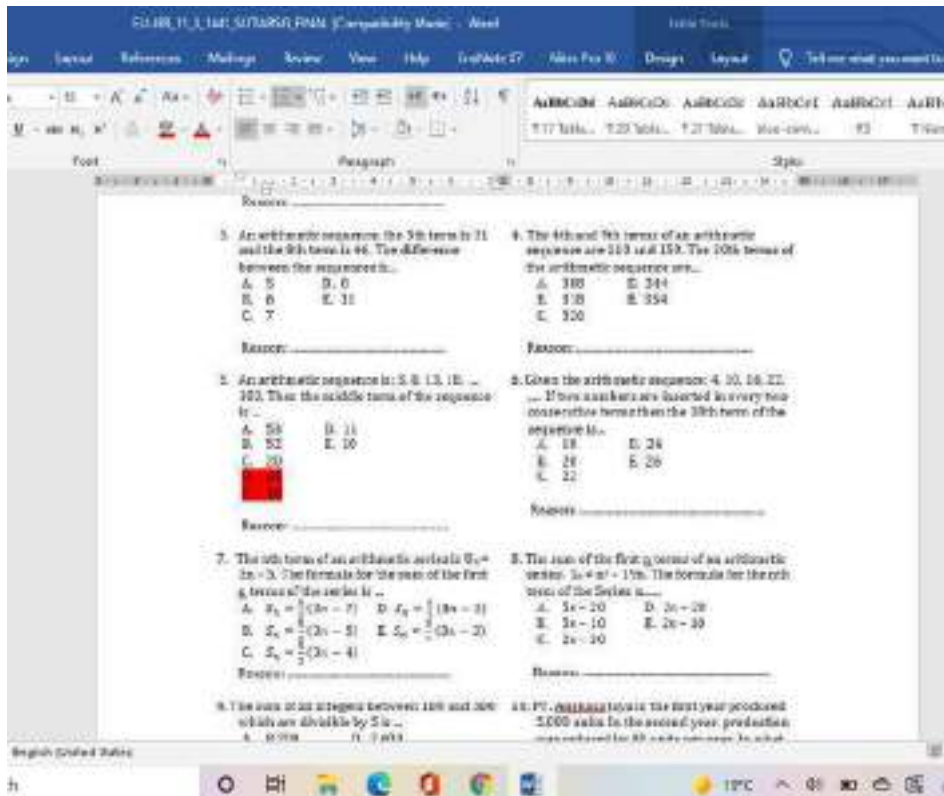
Sat, May 21, 2022 at 5:09 PM

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

I have checked again, there is still 1 typing error, namely: in appendix no 5 there is double writing on the answer choices D and E (highlighted in red), and I have deleted the answer choices.


Here I resubmit the article that I have revised (ok_final article).

Thank you for your kindness in revising my article.



Best regards,
Sugeng Sutiarso
University of Lampung

[Quoted text hidden]

 Ok_EU-JER_11_3_1441_SUTIARSO_FINAL.docx
1802K



European Journal of Educational Research

Volume 11, Issue 3, 1441 - 1462.

ISSN: 2165-8714

<https://www.eu-jer.com/>

Developing Assessment Instrument Using Polytomous Response in Mathematics

SugengSutiarso*

University of Lampung, INDONESIA

UndangRosidin

University of Lampung, INDONESIA

AanSulistiawan

Vocational School, INDONESIA

Received: November 25, 2021 • Revised: February 2, 2022 • Accepted: May 17, 2022

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used an open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) the open polytomous response test has a good category according to classical and modern theory. However, the discrimination power of test items in classical theory needs several revisions, (2) the assessment instrument using the polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses.

To cite this article: Sutiarso, S., Rosidin, U., & Sulistiawan, A. (2022). Developing assessment instrument using polytomous response in mathematics. *European Journal of Educational Research*, 11(3), 1441-1462. <https://doi.org/10.12973/eu-jer.11.3.1441>

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools was divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).



Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with specific rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over essay test are firstly, the test can be

* Corresponding author:

SugengSutiarso, University of Lampung, Sumantri Brojonegoro, Bandar Lampung, Indonesia. ✉ sugeng.sutiarso@fkip.unila.ac.id

© 2022 The Author(s). **Open Access**- This article is under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>)



conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weaknesses. The multiple-choice test is not able to portray the actual abilities of students, the answers of this test tend to be guessing games or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). A multiple-choice test with only two answer choices is called a "dichotomous test," and a multiple-choice test with more than two answer choices is called a "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or henceforth called as polytomous response test (Suwarto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as a test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), a test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into an open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response tests for students in college and senior or junior school. Students in college and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools are not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomous test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmalawati, 2020). Therefore, the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomous response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing the item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because, in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows: (1) Does

the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory? and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's (2013) model, with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

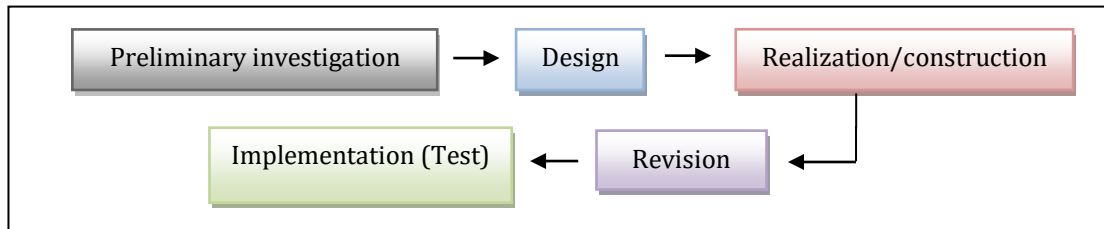


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the

indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. Questionnaire data analysis (qualitative analysis)

There are two sets of questionnaire data, namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically, for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. Test data analysis (empirical analysis)

After conducting the content validity test, the researcher conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to have good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program was used for classical theory

and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep) because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Correlation column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit's MNSQ value is 0.5 to 1.5 and the Pt-Measure Correlation is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay tests and 20% of teachers used multiple-choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

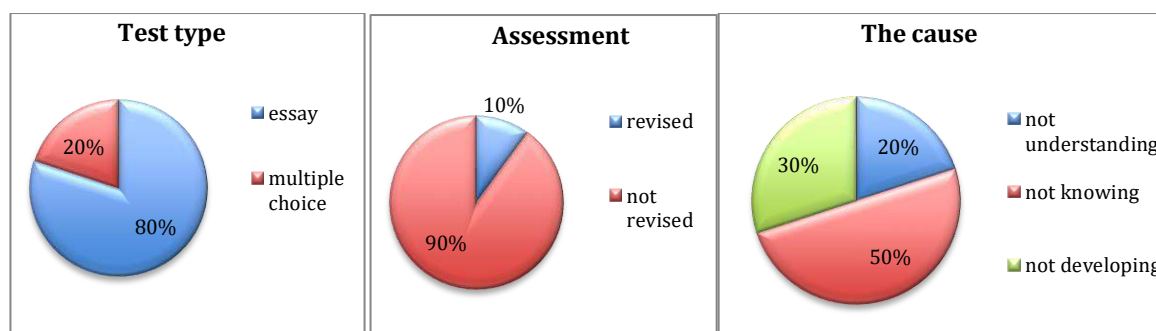


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	Good	0.196	Revised	21	0.482	Good	0.143	Revised
2	0.487	Good	0.179	Revised	22	0.535	Good	0.429	Good
3	0.528	Good	0.214	Revised	23	0.492	Good	0.250	Revised
4	0.540	Good	0.304	Good	24	0.438	Good	-0.071	Revised
5	0.489	Good	0.089	Revised	25	0.436	Good	-0.107	Revised
6	0.446	Good	-0.161	Revised	26	0.385	Good	-0.286	Revised
7	0.438	Good	-0.232	Revised	27	0.383	Good	-0.321	Revised
8	0.453	Good	-0.143	Revised	28	0.416	Good	-0.143	Revised
9	0.414	Good	-0.143	Revised	29	0.458	Good	-0.125	Revised
10	0.409	Good	-0.339	Revised	30	0.385	Good	-0.375	Revised
11	0.438	Good	-0.143	Revised	31	0.404	Good	-0.321	Revised
12	0.436	Good	-0.036	Revised	32	0.433	Good	-0.250	Revised
13	0.400	Good	-0.321	Revised	33	0.441	Good	0,036	Revised
14	0.450	Good	-0.036	Revised	34	0.424	Good	-0.268	Revised
15	0.462	Good	0.250	Revised	35	0.412	Good	-0.321	Revised
16	0.453	Good	-0.089	Revised	36	0.431	Good	-0.304	Revised
17	0.416	Good	-0.143	Revised	37	0.404	Good	-0.232	Revised
18	0.419	Good	-0.196	Revised	38	0.363	Good	-0.482	Revised
19	0.431	Good	-0.232	Revised	39	0.230	Good	-0.929	Revised
20	0.441	Good	-0.089	Revised	40	0.211	Good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

*Analysis of Test Data with Modern Theory**The Unidimensional Assumption Test*

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first-factor Eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first-factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

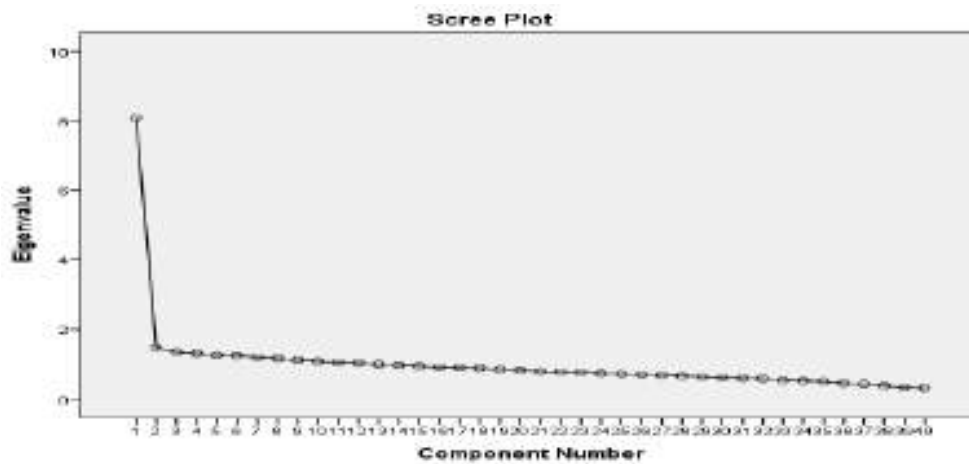


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Correlation is positive, then it can be said that the item fits the model (Sumintono&Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-0.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-0.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-0.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-0.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-0.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-0.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-0.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-0.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-0.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-0.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained were presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono&Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-0.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-0.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-0.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-0.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-0.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-0.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-0.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-0.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-0.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-0.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

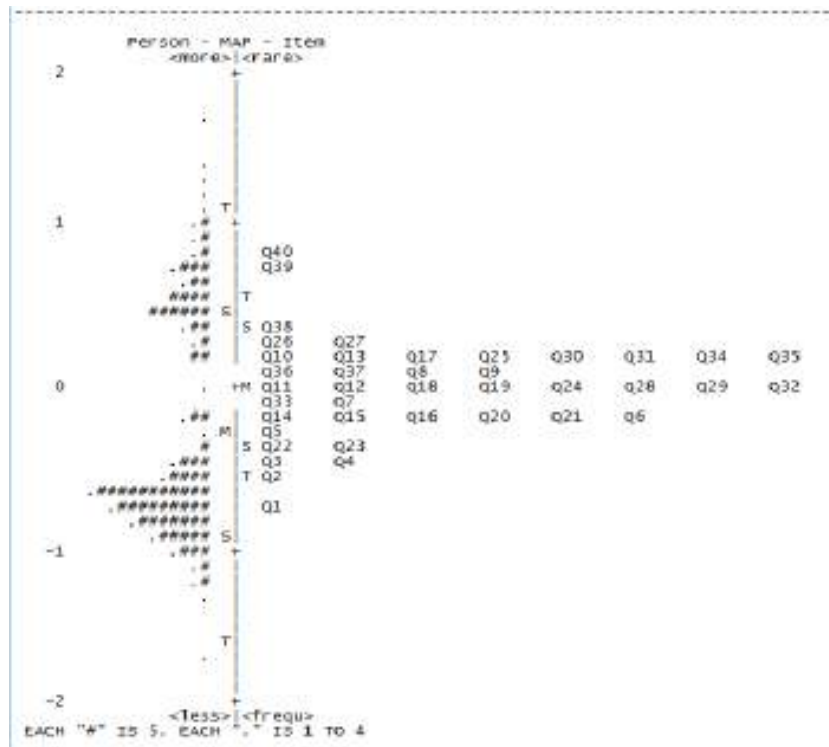


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (PT-Measure Correlation column). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

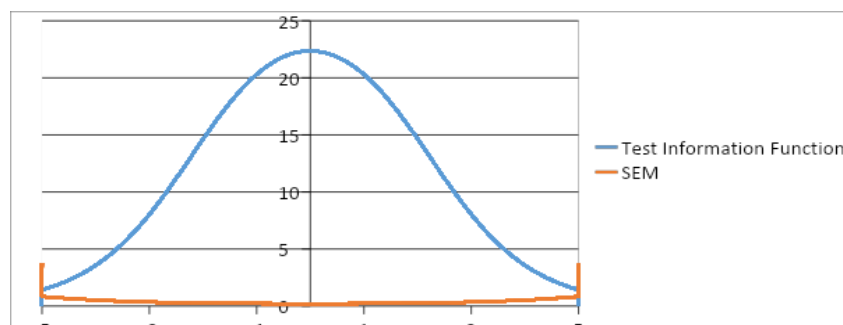


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in the vocational school was based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

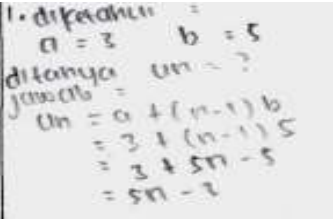
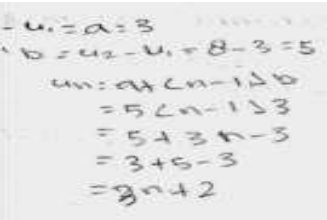
<p>Question 1:</p> <p>Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is ...</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	<p>Pattern 1:</p> 	<p>Pattern 2:</p> 
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term) and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

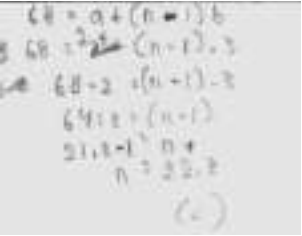
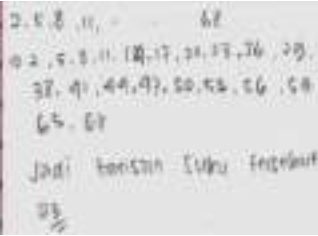
<p>Question 2:</p> <p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>	<p>Pattern 1</p> 	<p>Pattern 2</p> 
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

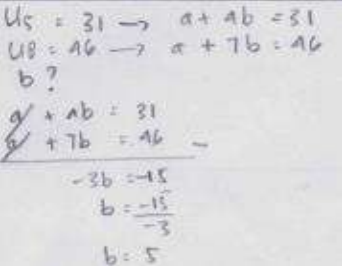
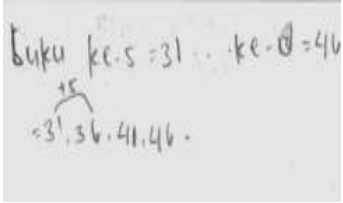
<p>Question 3:</p> <p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>	<p>Pattern 1</p> 	<p>Pattern 2</p> 
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------

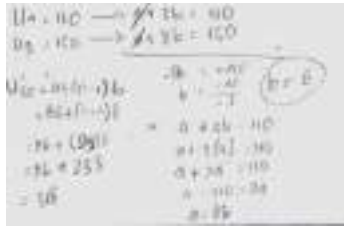
Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

Question 4:
 The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308
 B. 318
 C. 326
 D. 344
 E. 354

Reason:

Pattern 1



Pattern 2

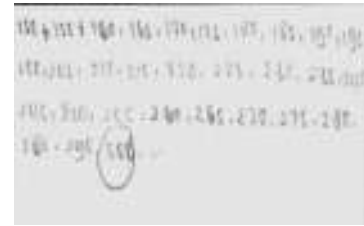


Figure 13. An Example of Student Answers in Item 4

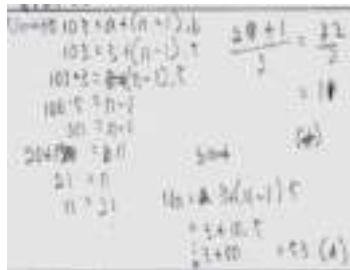
Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

Question 5:

An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
 Then the middle term of the sequence is ...308
 A. 53
 B. 52
 C. 20
 D. 11
 E. 10

Reason:

Answer Pattern 1



Answer Pattern 2



Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice test is easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score. Then, the result is multiplied by 10 to get a value range of 0–10 or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomus makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomus response test is more secondhand and complete than other assessment instruments, such as multiple-choice tests(Gierl et al., 2017) or essay tests(Putri et al., 2020).

Discussion

This research is development research aimed at developing a good mathematical assessment instrument using polytomous responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item

discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of the classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to the modern theory before revising or replacing the item.

Research on learning assessment with open response polytomous was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

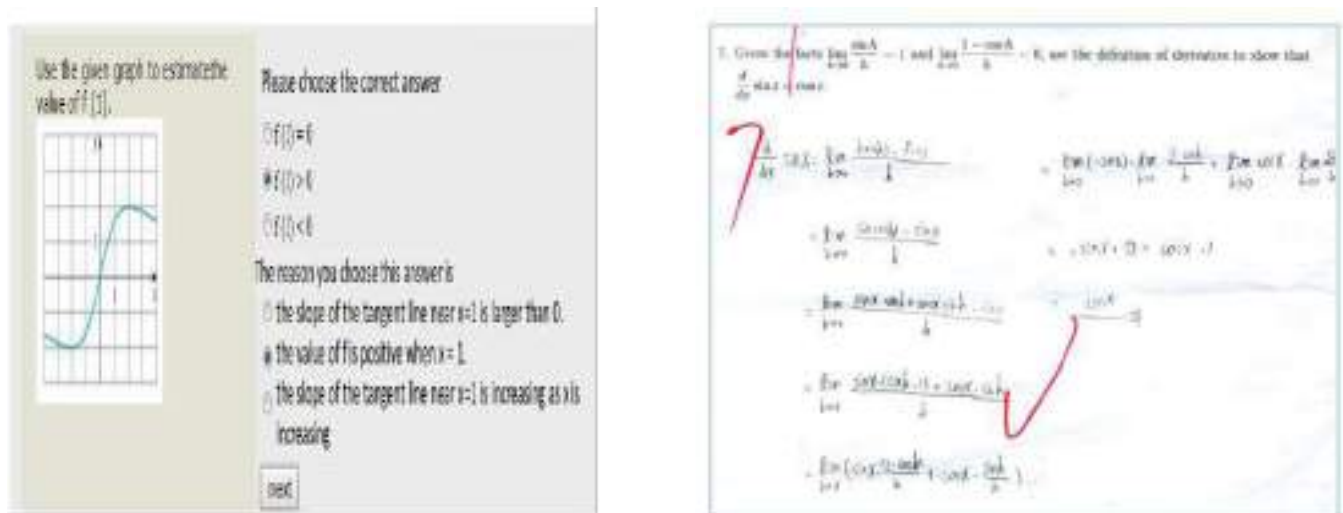


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous

Response Test

Another study on learning assessment with open response polychromous was conducted by Ayanwale(2021).Ayanwale's research compares two polytomous response test analysis methods namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

Other studies related to classical and modern theory were conducted bySarea (2018)and Saepuzaman et al.(2021). Sarea's research states that the response polytomous test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomous test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

- A. 308
- B. 318
- C. 326
- D. 344
- E. 354

Reason:

$$\begin{aligned}
 U_4 &= 110 \rightarrow a + 3b = 110 \\
 U_9 &= 150 \rightarrow a + 8b = 150 \\
 \hline
 -5b &= -40 \\
 b &= \frac{-40}{-5} \quad (b = 8) \\
 a + 3(8) &= 110 \\
 a + 24 &= 110 \\
 a &= 110 - 24 \\
 a &= 86 \\
 U_{30} &= a + (n-1)b \\
 &= 86 + (30-1)8 \\
 &= 86 + 233 \\
 &= 319
 \end{aligned}$$

(i)

(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, using a drawing or model, analogy, and formula. Syhlan(2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syhlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomus response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. Thus, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, it is suggested to conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

Authorship Contribution Statement

Sutiarso: Conceptualization, design, development of instruments, analysis, article writing, final approval. Rosidin: Design, development of instruments, analysis, editing. Sulistiawan: Development of instruments, collect data, analysis, editing.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskorandikotomi dan politomidalam teoriresponbutiruntuk pengembangan bank soalmatakuliah matematikadasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://bit.ly/39wv73P>
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campurandikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and generalized partial credit model]. *Jurnal Suluh Pendidikan*, 17(1), 32-37. <https://bit.ly/3yFQ6eO>
- Arikunto, S. (2012). *Dasar-dasarevaluasipendidikan* [Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://bit.ly/3Nob4mN>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://bit.ly/3lfARKW>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. <https://bit.ly/3MnA9hw>
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geoffrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. ResearchGate. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: Challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>

- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomis dan politomis pada tes prestasi belajar [Equalization of the dichotomous and polytomus mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: What, why, and how? In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <https://bit.ly/3liFMS8>
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <https://bit.ly/3sEcNMA>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://bit.ly/3Psrc8s>
- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://bit.ly/3wljq95>
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://bit.ly/3yIePzb>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standarisasi pendidikan dasar dan menengah* [Standards of content for primary and secondary education]. Indonesian Government publications service. <https://bit.ly/3FZHlhp>
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <https://bit.ly/3MzqkNm>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://bit.ly/37QCtyu>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://bit.ly/3yF4iF5>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya* [Item response theory and its application]. Nuha Medika. <https://bit.ly/39TFDFE>

- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163. <https://doi.org/10.30587/postulat.v1i2.2094>
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisika dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst: Journal of Physics Education and Its Application/Karst: Jurnal Pendidikan Fisika dan Terapannya*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budipekertingkatsekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://bit.ly/3FTtoOU7>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Teori tes klasik and respon [Characteristics of items: Classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <https://doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data peneliti dengan model rash dan Winstep* [Analysis of research data with rash and Winstep models]. Halaman Moeka Publishing. <https://bit.ly/3NlnGLE>
- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <https://bit.ly/3wjuXp9>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student:

Class/Department :

School :

Instructions: Mark (x) one of the correct answer choices, and give the reason

(use another piece of paper to write down your reason)

1. Given an arithmetic sequence: 3, 8, 13, 18, ... The formula for the n th term of the sequence is ...
 A. $U_n = 5n - 3$ C. $U_n = 4n - 1$
 B. $U_n = 5n - 2$ D. $U_n = 3n + 2$
 C. $U_n = 4n - 1$
 Reason:
2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...
 A. 12 D. 23
 B. 13 E. 24
 C. 22
 Reason:
3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
 A. 5 D. 8
 B. 6 E. 11
 C. 7
 Reason:
4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308 D. 344
 B. 318 E. 354
 C. 326
 Reason:
5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...
 A. 53 D. 11
 B. 52 E. 10
 C. 20
 D. 11
 E. 10
 Reason:
6. Given the arithmetic sequence: 4, 10, 16, 22, ... If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
 A. 18 D. 24
 B. 20 E. 26
 C. 22
 Reason:
7. The n th term of an arithmetic series is $U_n = 3n - 5$. The formula for the sum of the first n terms of the series is ...
 A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
 B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
 C. $S_n = \frac{n}{2}(3n - 4)$
 Reason:
8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the n th term of the Series is.....
 A. $5n - 20$ D. $2n - 20$
 B. $5n - 10$ E. $2n - 10$
 C. $2n - 30$
 Reason:
9. The sum of all integers between 100 and 300 which are divisible by 5 is ...
 A. 8.200 D. 7.600
 B. 8.000 E. 7.400
 C. 7.800
 Reason:
10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
 A. 24 D. 27
 B. 25 E. 28
 C. 26
 Reason:

11. The middle term of an arithmetic sequence is 25. If the difference is 4 and the 5th term is 21. Then the sum of all the terms in the sequence is ...

- A. 175 D. 295
- B. 189 E. 375
- C. 275

Reason:

12. A number of candies are distributed among five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces

- A. 60 D. 75
- B. 65 E. 80
- C. 70

Reason:

13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...

- A. 564 D. 45
- B. 276 E. 36
- C. 48

Reason:

14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...

- A. 9 D. 12
- B. 10 E. 13
- C. 11

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...

- A. 32 D. 256
- B. 64 E. 512
- C. 128

Reason:

16. The value of the middle term of the geometric sequence: 6, 3, ..., $3/512$ is ...

- A. $\frac{1}{16}$ D. $\frac{4}{16}$
- B. $\frac{2}{16}$ E. $\frac{5}{16}$
- C. $\frac{3}{16}$

Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm

- A. 18 D. 35
- B. 24 E. 40,5
- C. 27,5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...

- A. $\frac{3}{4}$ D. $-\frac{1}{2}$
- B. $\frac{1}{4}$ E. $-\frac{3}{4}$
- C. $\frac{4}{3}$

Reason:

19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is... m

- A. 60 D. 90
- B. 70 E. 100
- C. 80

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.

The value of $3a + b$ is ...

- A. 8 D. 14
- B. 10 E. 20
- C. 12

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$. If $K = L$, then cis ...

- A. 12 D. 15
- B. 13 E. 16
- C. 14

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$. Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$
- D. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ is

- ...
- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ is...

A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...

A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ...

A. 0 D. 2
 B. 1 E. 4
 C. 2

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$. Inverse matrix $(AB)^{-1}$ = ...

- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$

Reason:

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.

If $A + B = C$, then $x + y = \dots$

- A. -5 D. 3
 B. -1 E. 5
 C. 1

Reason:

26. If $\begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$ then $A(B - C) = \dots$

- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix. Value of x that satisfies is ...

- A. -5 D. 3
 B. -4 E. 4
 C. -3

Reason:

30. Transpose matrix P is P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ...

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$

Reason:

32. The roots of the quadratic equation $3x^2 - 4x + 3 = 0$ are ...

- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$

Reason:

33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...
 A. -2 and $\frac{5}{6}$ D. -2 and $-\frac{6}{5}$
 B. 2 and $-\frac{5}{6}$ E. -2 and $\frac{6}{5}$
 C. 2 and $\frac{6}{5}$

Reason:

35. The roots of the quadratic equation: $2x^2 - 3x - 9 = 0$ are x_1 and x_2 . Value of $x_1^2 + x_2^2$ is ...
 A. $11\frac{1}{4}$ D. $-6\frac{3}{4}$
 B. $6\frac{3}{4}$ E. $-11\frac{1}{4}$
 C. $2\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...
 A. $y = x^2 - 2x + 1$
 B. $y = x^2 - 2x + 3$
 C. $y = x^2 + 2x - 1$
 D. $y = x^2 + 2x + 1$
 E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0,-4)$ then the value of $f(7)$ is ...
 A. -16 D. -19
 B. -17 E. -20
 C. -18

Reason:

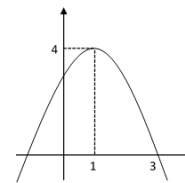
34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...
 A. -4 D. 2
 B. -2 E. 4
 C. 0

Reason:

36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α and β . The quadratic equation $(\alpha - 2)$ and $(\beta - 2)$ is ...
 a. $x^2 + 6x + 5 = 0$
 b. $x^2 + 6x + 7 = 0$
 c. $x^2 + 6x + 11 = 0$
 d. $x^2 - 2x + 3 = 0$
 e. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation? ...
 A. $y = x^2 + 2x + 3$
 B. $y = x^2 - 2x - 3$
 C. $y = -x^2 + 2x - 3$
 D. $y = -x^2 - 2x + 3$
 E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...
 A. $y = -x^2 + 2x - 3$ D. $y = -x^2 - 2x - 5$
 B. $y = -x^2 + 2x + 3$ E. $y = -x^2 - 2x + 5$
 C. $y = -x^2 - 2x + 3$

Reason:



European Journal of Educational Research

Volume 11, Issue 3, 1441 - 1462.

ISSN: 2165-8714

<https://www.eu-jer.com/>

Developing Assessment Instrument Using Polytomous Response in Mathematics

SugengSutiarso*

University of Lampung, INDONESIA

UndangRosidin

University of Lampung, INDONESIA

AanSulistiawan

Vocational School, INDONESIA

Received: November 25, 2021 • Revised: February 2, 2022 • Accepted: May 17, 2022

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used an open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) the open polytomous response test has a good category according to classical and modern theory. However, the discrimination power of test items in classical theory needs several revisions, (2) the assessment instrument using the polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses.

To cite this article: Sutiarso, S., Rosidin, U., & Sulistiawan, A. (2022). Developing assessment instrument using polytomous response in mathematics. *European Journal of Educational Research*, 11(3), 1441-1462. <https://doi.org/10.12973/eu-jer.11.3.1441>

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools was divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).

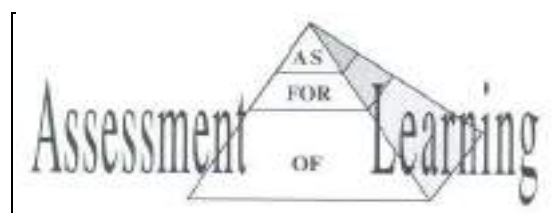


Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with specific rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over essay test are firstly, the test can be

* Corresponding author:

SugengSutiarso, University of Lampung, Sumantri Brojonegoro, Bandar Lampung, Indonesia. ✉ sugeng.sutiarso@fkip.unila.ac.id

© 2022 The Author(s). **Open Access**- This article is under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>)



conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weaknesses. The multiple-choice test is not able to portray the actual abilities of students, the answers of this test tend to be guessing games or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). A multiple-choice test with only two answer choices is called a "dichotomous test," and a multiple-choice test with more than two answer choices is called a "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or henceforth called as polytomous response test (Suwarto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as a test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), a test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into an open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response tests for students in college and senior or junior school. Students in college and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools are not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomous test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmalawati, 2020). Therefore, the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomous response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing the item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because, in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows: (1) Does

the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory? and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's (2013) model, with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

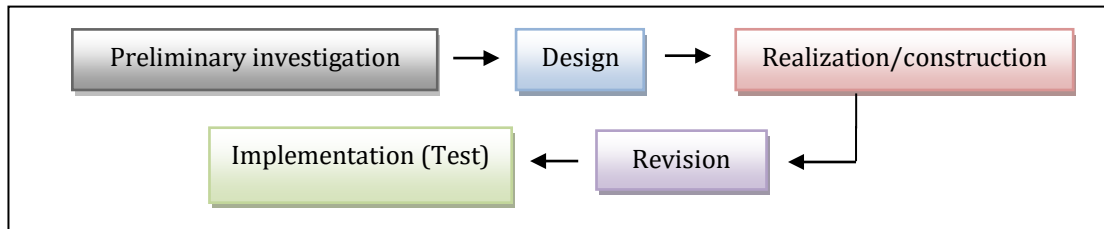


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the

indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. Questionnaire data analysis (qualitative analysis)

There are two sets of questionnaire data, namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically, for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. Test data analysis (empirical analysis)

After conducting the content validity test, the researcher conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to have good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program was used for classical theory

and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1 Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2 Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep) because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Correlation column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit's MNSQ value is 0.5 to 1.5 and the Pt-Measure Correlation is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay tests and 20% of teachers used multiple-choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

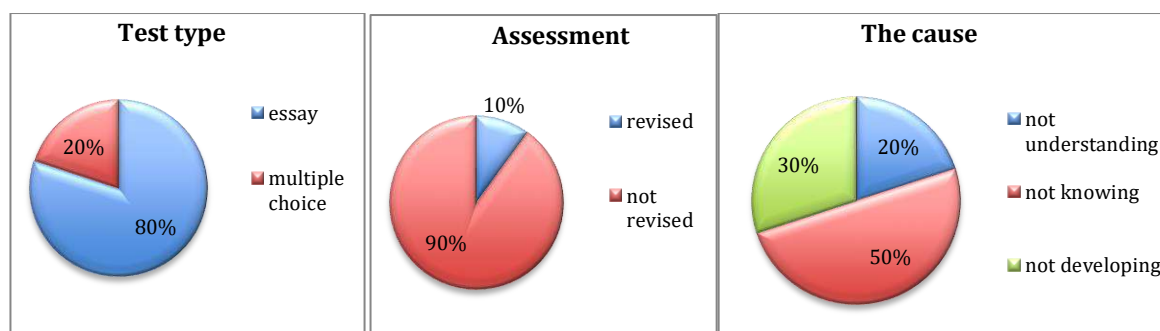


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	Good	0.196	Revised	21	0.482	Good	0.143	Revised
2	0.487	Good	0.179	Revised	22	0.535	Good	0.429	Good
3	0.528	Good	0.214	Revised	23	0.492	Good	0.250	Revised
4	0.540	Good	0.304	Good	24	0.438	Good	-0.071	Revised
5	0.489	Good	0.089	Revised	25	0.436	Good	-0.107	Revised
6	0.446	Good	-0.161	Revised	26	0.385	Good	-0.286	Revised
7	0.438	Good	-0.232	Revised	27	0.383	Good	-0.321	Revised
8	0.453	Good	-0.143	Revised	28	0.416	Good	-0.143	Revised
9	0.414	Good	-0.143	Revised	29	0.458	Good	-0.125	Revised
10	0.409	Good	-0.339	Revised	30	0.385	Good	-0.375	Revised
11	0.438	Good	-0.143	Revised	31	0.404	Good	-0.321	Revised
12	0.436	Good	-0.036	Revised	32	0.433	Good	-0.250	Revised
13	0.400	Good	-0.321	Revised	33	0.441	Good	0,036	Revised
14	0.450	Good	-0.036	Revised	34	0.424	Good	-0.268	Revised
15	0.462	Good	0.250	Revised	35	0.412	Good	-0.321	Revised
16	0.453	Good	-0.089	Revised	36	0.431	Good	-0.304	Revised
17	0.416	Good	-0.143	Revised	37	0.404	Good	-0.232	Revised
18	0.419	Good	-0.196	Revised	38	0.363	Good	-0.482	Revised
19	0.431	Good	-0.232	Revised	39	0.230	Good	-0.929	Revised
20	0.441	Good	-0.089	Revised	40	0.211	Good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

*Analysis of Test Data with Modern Theory**The Unidimensional Assumption Test*

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first-factor Eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first-factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

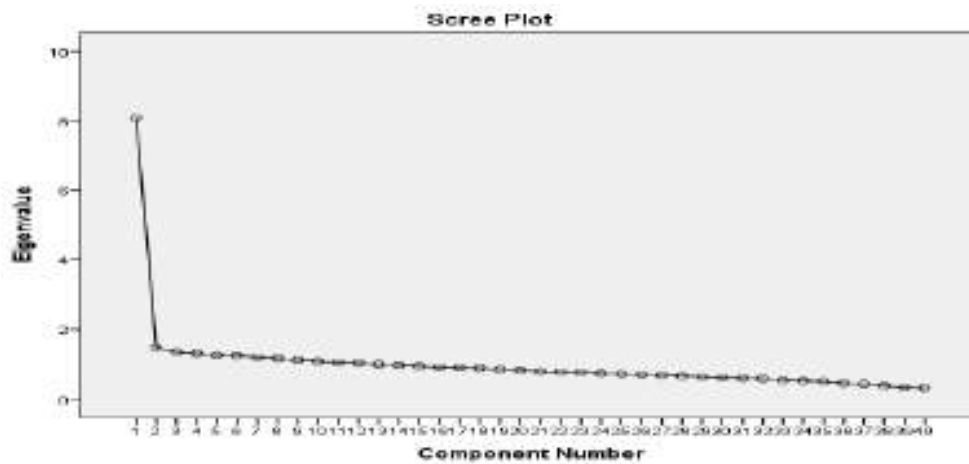


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Correlation is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-0.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-0.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-0.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-0.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-0.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-0.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-0.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.6	Q22
23	1037	413	-0.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-0.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-0.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained were presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-0.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-0.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-0.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-0.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-0.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-0.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-0.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.6	Q22
23	1037	413	-0.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-0.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-0.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

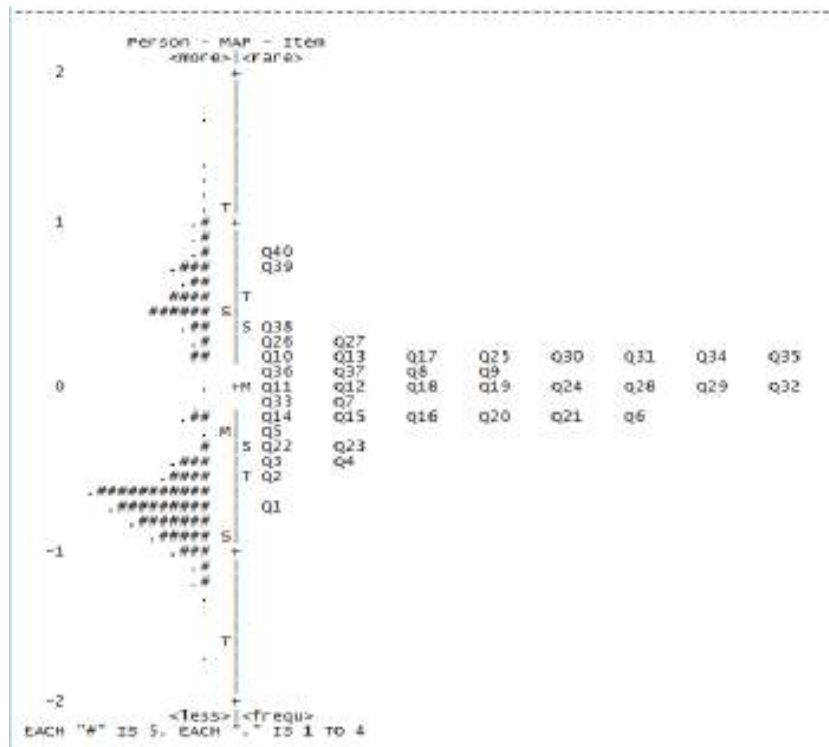


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (PT-Measure Correlation column). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

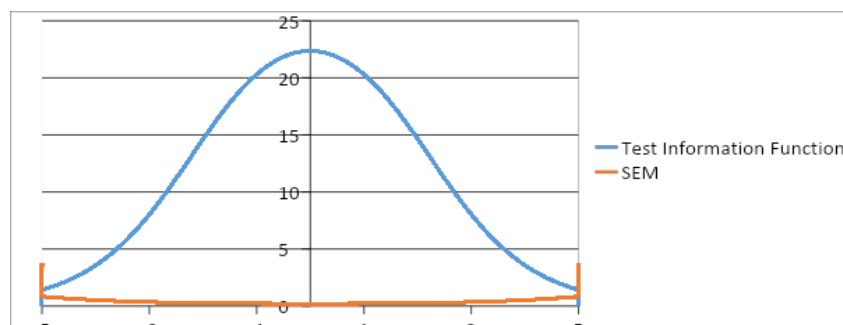


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in the vocational school was based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

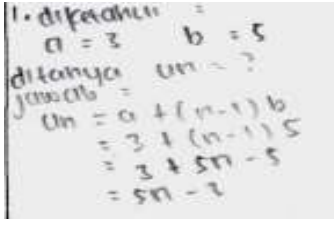
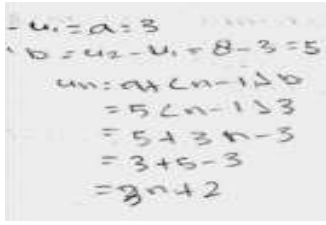
<p style="text-align: center;">Question 1:</p> <p>Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is ...</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	<p style="text-align: center;">Pattern 1:</p> 	<p style="text-align: center;">Pattern 2:</p> 
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term) and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

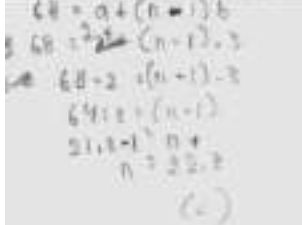
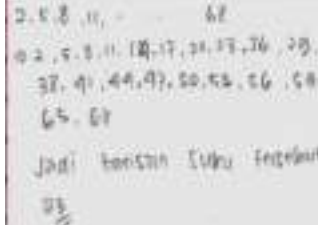
<p style="text-align: center;">Question 2:</p> <p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>	<p style="text-align: center;">Pattern 1</p> 	<p style="text-align: center;">Pattern 2</p> 
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

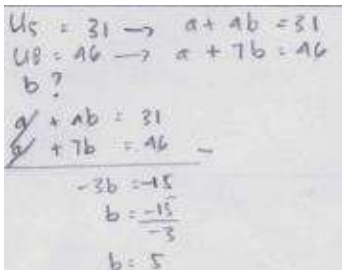
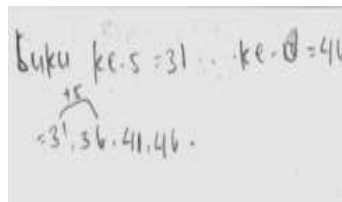
<p style="text-align: center;">Question 3:</p> <p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>	<p style="text-align: center;">Pattern 1</p> 	<p style="text-align: center;">Pattern 2</p> 
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------

Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

Question 4:
 The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308
 B. 318
 C. 326
 D. 344
 E. 354

Reason:

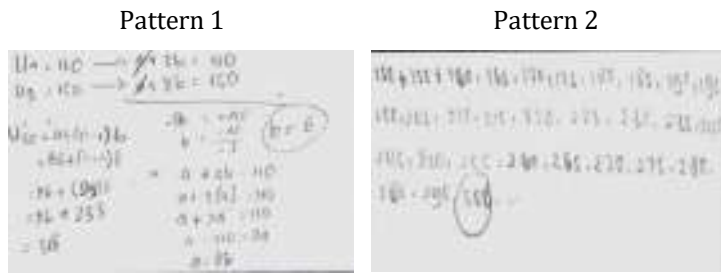


Figure 13. An Example of Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

Question 5:
 An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
 Then the middle term of the sequence is ...308
 A. 53
 B. 52
 C. 20
 D. 11
 E. 10

Reason:

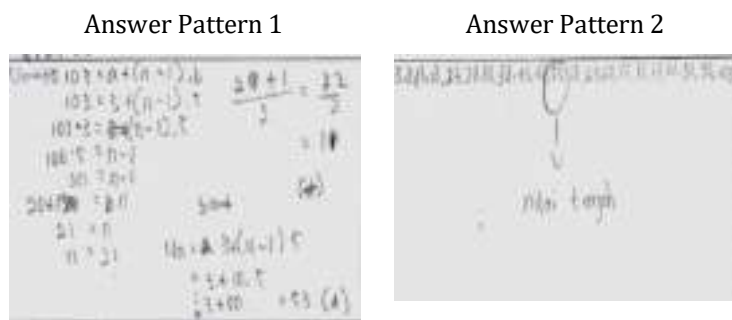


Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice test is easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score. Then, the result is multiplied by 10 to get a value range of 0–10 or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomous makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomous response test is more secondhand and complete than other assessment instruments, such as multiple-choice tests(Gierl et al., 2017) or essay tests(Putri et al., 2020).

Discussion

This research is development research aimed at developing a good mathematical assessment instrument using polytomous responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item

discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of the classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to the modern theory before revising or replacing the item.

Research on learning assessment with open response polytomous was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

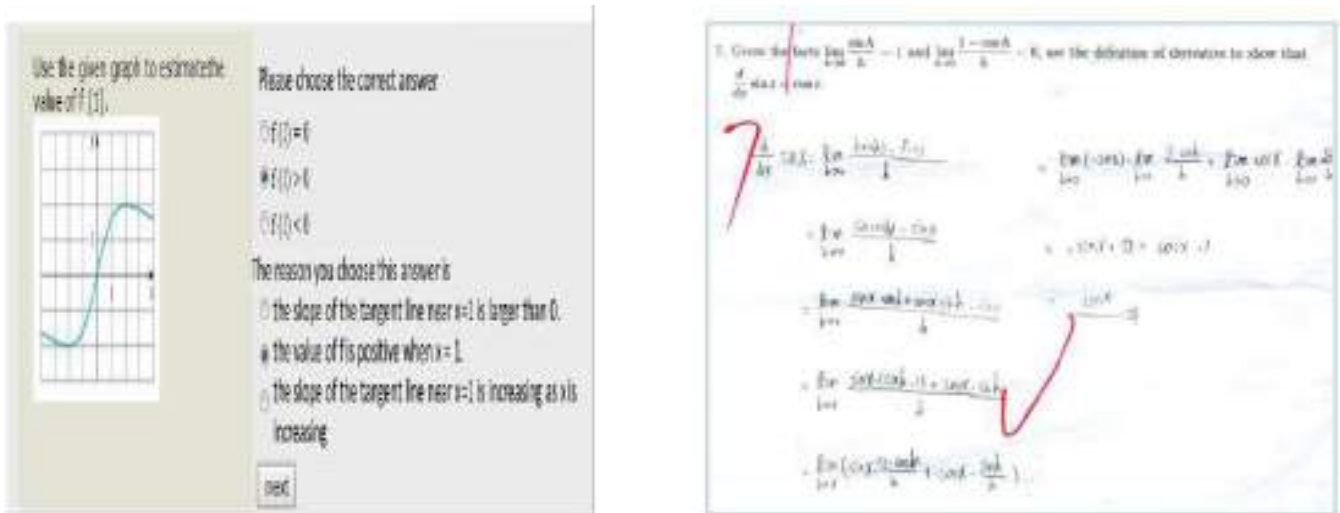


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous

Response Test

Another study on learning assessment with open response polychromous was conducted by Ayanwale(2021).Ayanwale's research compares two polytomous response test analysis methods namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

Other studies related to classical and modern theory were conducted bySarea (2018)and Saepuzaman et al.(2021). Sarea's research states that the response polytomous test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomous test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulasand (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

- A. 308
B. 318
C. 326
D. 344
E. 354

Reason:

$$\begin{aligned}
 U_4 &= 110 \rightarrow a + 3b = 110 \\
 U_9 &= 150 \rightarrow a + 8b = 150 \\
 \hline
 & -5b = -40 \quad (b = 8) \\
 & b = \frac{-40}{-5} \\
 & = 8 \\
 & a + 3(8) = 110 \\
 & a + 24 = 110 \\
 & a = 110 - 24 \\
 & a = 86 \\
 U_{30} &= a + (n-1)b \\
 &= 86 + (30-1)8 \\
 &= 86 + (29)8 \\
 &= 86 + 233 \\
 &= 319
 \end{aligned}$$

(i)

(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, using a drawing or model, analogy, and formula. Syhlan(2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syhlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomous response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. Thus, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, it is suggested to conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

Authorship Contribution Statement

Sutiarso: Conceptualization, design, development of instruments, analysis, article writing, final approval. Rosidin: Design, development of instruments, analysis, editing. Sulistiawan: Development of instruments, collect data, analysis, editing.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://bit.ly/39wv73P>
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomis dan politomis generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and polytomous generalized partial credit model]. *JurnalSuluh Pendidikan*, 17(1), 32-37. <https://bit.ly/3yFQ6eO>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://bit.ly/3Nob4mN>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://bit.ly/3lfARkW>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. <https://bit.ly/3MnA9hw>
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geoffrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. ResearchGate. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: Challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>

- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar [Equalization of the dichotomous and polytomous mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: What, why, and how? In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <https://bit.ly/3liFMS8>
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <https://bit.ly/3sEcNMA>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://bit.ly/3Psrc8s>
- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://bit.ly/3wljq95>
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://bit.ly/3yIePzb>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Standards of content for primary and secondary education]. Indonesian Government publicationservice. <https://bit.ly/3FZHlhp>
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <https://bit.ly/3MzqkNm>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://bit.ly/37QCtyu>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://bit.ly/3yF4iF5>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya* [Item response theory and its application]. NuhaMedika. <https://bit.ly/39TFDFI>

- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163. <https://doi.org/10.30587/postulat.v1i2.2094>
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst: Journal of Physics Education and Its Application/Karst: Jurnal Pendidikan Fisika dan Terapannya*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://bit.ly/3FToOU7>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Teori tes klasik and respon [Characteristics of items: Classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <https://doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan Winstep* [Analysis of research data with rash and Winstep models]. Halaman Moeka Publishing. <https://bit.ly/3NlnGLc>
- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <https://bit.ly/3wjuXp9>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :

Class/Department :

School :

Instructions:Mark (x) one of the correct answer choices, and give the reason

(use another piece of paper to write down your reason)

1. Given an arithmetic sequence: 3, 8, 13, 18, ... The formula for the nth term of the sequence is ...
 A. $U_n = 5n - 3$ D. $U_n = 4n - 1$
 B. $U_n = 5n - 2$ E. $U_n = 3n + 2$
 C. $U_n = 4n - 1$
 Reason:
2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...
 A. 12 D. 23
 B. 13 E. 24
 C. 22
 Reason:
3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
 A. 5 D. 8
 B. 6 E. 11
 C. 7
 Reason:
4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308 D. 344
 B. 318 E. 354
 C. 326
 Reason:
5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...
 A. 53 D. 11
 B. 52 E. 10
 C. 20
 Reason:
6. Given the arithmetic sequence: 4, 10, 16, 22, ... If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
 A. 18 D. 24
 B. 20 E. 26
 C. 22
 Reason:
7. The nth term of an arithmetic series is $U_n = 3n - 5$. The formula for the sum of the first n terms of the series is ...
 A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
 B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
 C. $S_n = \frac{n}{2}(3n - 4)$
 Reason:
8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....
 A. $5n - 20$ D. $2n - 20$
 B. $5n - 10$ E. $2n - 10$
 C. $2n - 30$
 Reason:
9. The sum of all integers between 100 and 300 which are divisible by 5 is ...
 A. 8,200 D. 7,600
 B. 8,000 E. 7,400
 C. 7,800
 Reason:
10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
 A. 24 D. 27
 B. 25 E. 28
 C. 26
 Reason:
11. The middle term of an arithmetic sequence
12. A number of candies are distributed among

is 25. If the difference is 4 and the 5th term is 21. Then the sum of all the terms in the sequence is ...

- A. 175 D. 295
- B. 189 E. 375
- C. 275

Reason:

five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces

- A. 60 D. 75
- B. 65 E. 80
- C. 70

Reason:

13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...

- A. 564 D. 45
- B. 276 E. 36
- C. 48

Reason:

14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...

- A. 9 D. 12
- B. 10 E. 13
- C. 11

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...

- A. 32 D. 256
- B. 64 E. 512
- C. 128

Reason:

16. The value of the middle term of the geometric sequence:

6, 3, ..., $3/512$ is ...

- A. $\frac{1}{16}$ D. $\frac{4}{16}$
- B. $\frac{2}{16}$ E. $\frac{5}{16}$
- C. $\frac{3}{16}$

Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm

- A. 18 D. 35
- B. 24 E. 40.5
- C. 27.5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...

- A. $\frac{3}{4}$ D. $-\frac{1}{2}$
- B. $\frac{1}{4}$ E. $-\frac{3}{4}$
- C. $\frac{1}{3}$

Reason:

19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is... m

- A. 60 D. 90
- B. 70 E. 100
- C. 80

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.

The value of $3a + b$ is ...

- A. 8 D. 14
- B. 10 E. 20
- C. 12

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$. If $K = L$, then cis ...

- A. 12 D. 15
- B. 13 E. 16
- C. 14

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$. Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$
- D. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$

is ...

- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ is...

- A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$

is ...

- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then

matrix determinant A is ...

- A. 0 D. 2
 B. 1 E. 4
 C. 2

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$.

Inverse matrix $(AB)^{-1}$ = ...

- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$

Reason:

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$,
 and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.

If $A + B = C$, then $x + y = \dots$

- A. -5 D. 3
 B. -1 E. 5
 C. 1

Reason:

26. If $A = \begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and

$C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$ then $A(B - C) = \dots$

- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix.

Value of x that satisfies is ...

- A. -5 D. 3
 B. -4 E. 4
 C. -3

Reason:

30. Transpose matrix P is P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then

matrix $(P^T)^{-1}$ is ...

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$

Reason:

32. The roots of the quadratic equation $3x^2 - 4x + 3 = 0$ are ...

- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$

Reason:

33. The roots of the quadratic equation $5x^2 + 4x - 12 = 0$ are ...
 A. -2 and $-\frac{5}{6}$ D. -2 and $-\frac{6}{5}$
 B. 2 and $-\frac{5}{6}$ E. -2 and $\frac{6}{5}$
 C. 2 and $\frac{6}{5}$

Reason:

35. The roots of the quadratic equation: $2x^2 - 3x - 9 = 0$ are x_1 and x_2 . Value of $x_1^2 + x_2^2$ is ...
 A. $11\frac{1}{4}$ D. $-6\frac{3}{4}$
 B. $6\frac{3}{4}$ E. $-11\frac{1}{4}$
 C. $2\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...
 A. $y = x^2 - 2x + 1$
 B. $y = x^2 - 2x + 3$
 C. $y = x^2 + 2x - 1$
 D. $y = x^2 + 2x + 1$
 E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0,-4)$ then the value of $f(7)$ is ...
 A. -16 D. -19
 B. -17 E. -20
 C. -18

Reason:

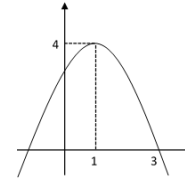
34. The roots of the equation $x^2 - 2x - 3 = 0$ are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...
 A. -4 D. 2
 B. -2 E. 4
 C. 0

Reason:

36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α and β . The quadratic equation $(\alpha - 2)$ and $(\beta - 2)$ is ...
 A. $x^2 + 6x + 5 = 0$
 B. $x^2 + 6x + 7 = 0$
 C. $x^2 + 6x + 11 = 0$
 D. $x^2 - 2x + 3 = 0$
 E. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation? ...
 A. $y = x^2 + 2x + 3$
 B. $y = x^2 - 2x - 3$
 C. $y = -x^2 + 2x - 3$
 D. $y = -x^2 - 2x + 3$
 E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...
 A. $y = -x^2 + 2x - 3$ D. $y = -x^2 - 2x - 5$
 B. $y = -x^2 + 2x + 3$ E. $y = -x^2 - 2x + 5$
 C. $y = -x^2 - 2x + 3$

Reason:



SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>

2nd Final Paper (ID#21112502244011)

2 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>

Sat, May 21, 2022 at 7:58 PM

Dear Dr. Sutiarso,

Thank you for your email. We have updated your paper.

Please find the attached 2nd finalized paper will be published.

We will publish it as online first soon.

The official publication date of your paper is July 15, 2022.

Best regards,
Ahmet Savas Ph.D.
Editor- European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

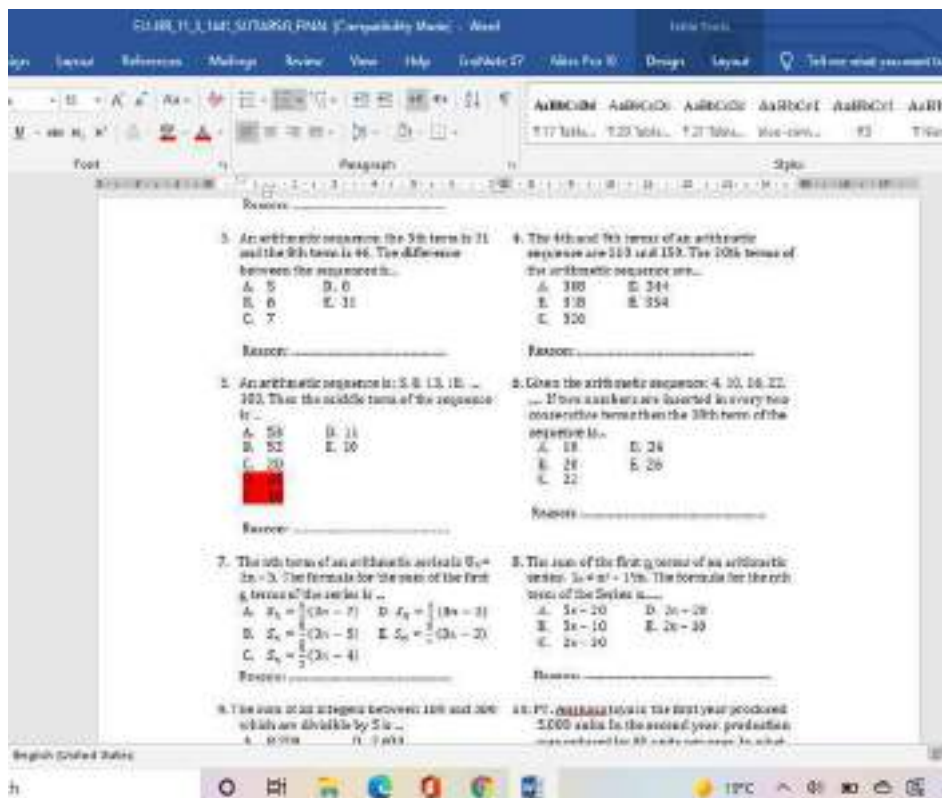
On 5/21/2022 1:09 PM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

I have checked again, there is still 1 typing error, namely: in appendix no 5 there is double writing on the answer choices D and E (highlighted in red), and I have deleted the answer choices.

Here I resubmit the article that I have revised (ok_final article).

Thank you for your kindness in revising my article.



Best regards,
Sugeng Sutiarmo
University of Lampung

On Sat, May 21, 2022 at 3:45 AM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarmo,

Thank you for your email. We have updated your paper. We have corrected a few mistakes.

Please find the attached finalized paper will be published.

We will publish it as online first soon.

The official publication date of your paper is July 15, 2022.

Best regards,
Ahmet Savas Ph.D.
Editor- European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 5/19/2022 5:14 PM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

I checked the language and references of my paper, and there are some edited parts of the appendix (highlighted in green), namely: (1) the word "and" (previously, the word "dan" in Indonesian) and (2) the word "is" (previously, the word "adalah" in Indonesian). Here is my final paper attached.

Best regards,
Sugeng Sutiarmo
Lampung University

On Tue, May 17, 2022 at 6:12 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarmo,

Please see the attached galley proof of your paper (ID#2103030737) (word file). Please highlight in green for your edited parts.

By the way,

- 1- Please check the language of your paper as a proofreading lastly.
- 2- Please check all references regarding with attached citation guide for APA 7 style. (Please see the citation guide page in our web site: <https://www.eujem.com/citation-guide>)

We ask you to check it please. Please edit at word file and resend it to me please in 2 days.

We are looking forward to getting your final paper by **May 19, 2022**.

Best regards,
Ahmet Savas Ph.D.
Editor, European Journal of Educational Management

<http://www.eujem.com>

editor@eujem.com

On 5/10/2022 1:24 PM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

Here I attach a revision of your proofreading suggestion.

Best regards,
Sugeng Sutiarto
Lampung University

On Fri, May 6, 2022 at 7:39 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarto,

Thank you for your kind email. Please see the attached file as the proofreading of your paper.

We are preparing the galley proof of your paper.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 4/23/2022 4:36 AM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

Here, I attach the copyright transfer agreement.

Best regards,
Sugeng Sutiarto
Lampung University

On Fri, Apr 22, 2022 at 5:24 PM European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sugeng Sutiarto,

We have received your payment about your paper entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" ID#21112502244011. Thanks.


We kindly ask from you to sign the copyright transfer agreement for your paper. After all author(s) signed, please scan and send via email to me **as soon as possible**. Please download the pdf file of this agreement from this link : <https://eu-jer.com/EU-JER-copyright-transfer-agreement.pdf> You can use e-signature, if you have. Also you can use your mobil phone as a scanner. If the other author live in another city, he/she sign the paper and send this paper via email. Than you can sign on this paper.

We are preparing the galley proof of your paper. We will send it to you in order to check before publication. The preparing of galley proofs may take some time because of our intensity. Thank you for your patience.

We are looking forward to getting copyright transfer agreement.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

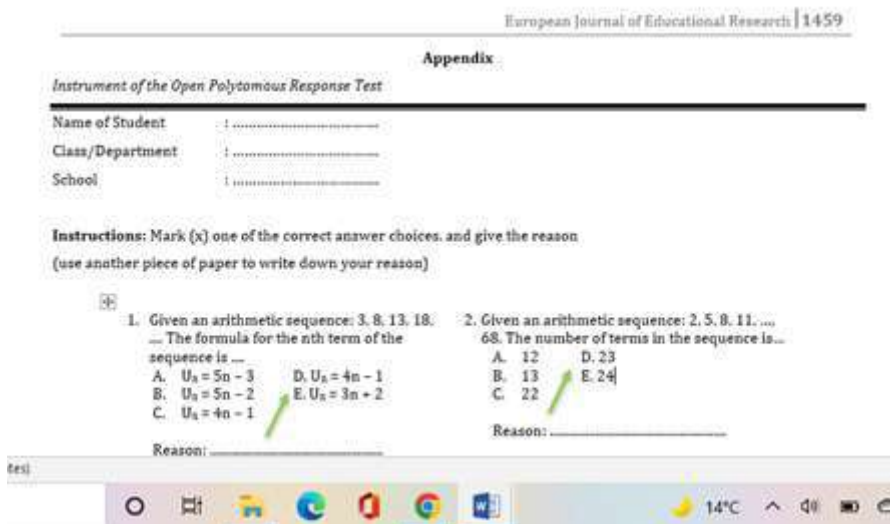
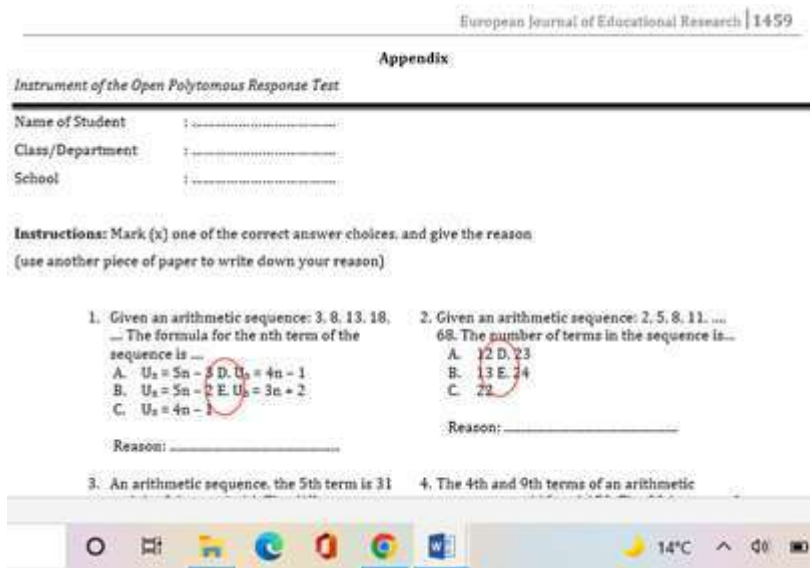
 EU-JER_11_3_1441_SUTIARSO_FINAL2.docx
1806K

SUGENG SUTIARSO <sugeng.sutiarso@fkip.unila.ac.id>
To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Sun, May 22, 2022 at 10:04 AM

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research.

I have checked that there are no more errors in the article substantially. However, I found a few writing errors in the answer choices (in the appendix), namely there is no space between the answer choices (as shown below).



I revised it, and here I resubmit the corrected article.
Hopefully, this article is final (no more mistakes).

Thank you for your kindness and patience in revising my article.

Best regards,
Sugeng Sutiarso
University of Lampung

[Quoted text hidden]



European Journal of Educational Research

Volume 11, Issue 3, 1441 - 1462.

ISSN: 2165-8714

<https://www.eu-jer.com/>

Developing Assessment Instrument Using Polytomous Response in Mathematics

SugengSutiarso*

University of Lampung, INDONESIA

UndangRosidin

University of Lampung, INDONESIA

AanSulistiawan

Vocational School, INDONESIA

Received: November25, 2021 • Revised: February2, 2022 • Accepted: May 17, 2022

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used an open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) the open polytomous response test has a good category according to classical and modern theory. However, the discrimination power of test items in classical theory needs several revisions, (2) the assessment instrument using the polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses.

To cite this article: Sutiarso, S., Rosidin, U., & Sulistiawan, A. (2022). Developing assessment instrument using polytomous response in mathematics. *European Journal of Educational Research*, 11(3), 1441-1462. <https://doi.org/10.12973/eu-jer.11.3.1441>

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools was divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).



Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with specific rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over essay test are firstly, the test can be

* Corresponding author:

SugengSutiarso, University of Lampung, Sumantri Brojonegoro, Bandar Lampung, Indonesia. ✉ sugeng.sutiarso@fkip.unila.ac.id

© 2022 The Author(s). **Open Access**- This article is under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>)



conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weaknesses. The multiple-choice test is not able to portray the actual abilities of students, the answers of this test tend to be guessing games or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). A multiple-choice test with only two answer choices is called a "dichotomous test," and a multiple-choice test with more than two answer choices is called a "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or henceforth called as polytomous response test (Suwarto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as a test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), a test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into an open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response tests for students in college and senior or junior school. Students in college and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools are not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomous test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmalawati, 2020). Therefore, the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomous response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing the item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because, in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows: (1) Does

the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory? and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's (2013) model, with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

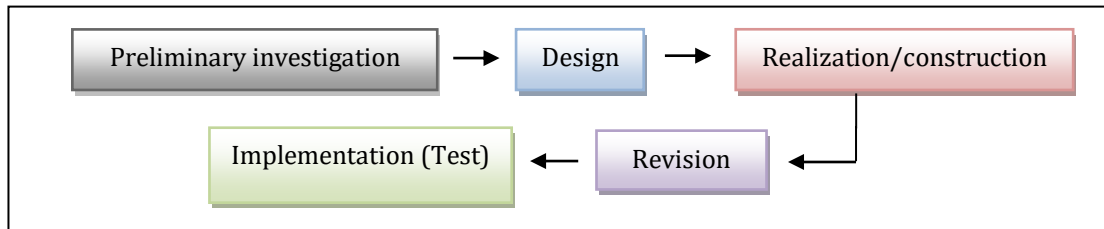


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the

indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. Questionnaire data analysis (qualitative analysis)

There are two sets of questionnaire data, namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically, for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. Test data analysis (empirical analysis)

After conducting the content validity test, the researcher conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to have good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program was used for classical theory

and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1. Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2. Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep) because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Correlation column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit's MNSQ value is 0.5 to 1.5 and the Pt-Measure Correlation is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay tests and 20% of teachers used multiple-choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

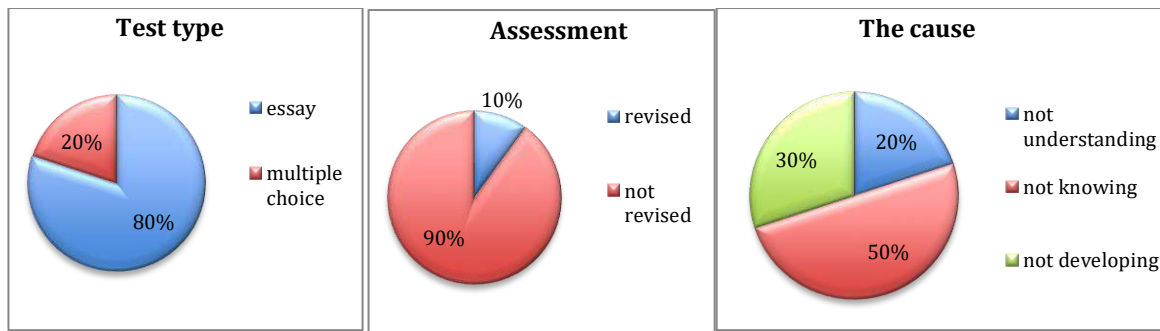


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	Good	0.196	Revised	21	0.482	Good	0.143	Revised
2	0.487	Good	0.179	Revised	22	0.535	Good	0.429	Good
3	0.528	Good	0.214	Revised	23	0.492	Good	0.250	Revised
4	0.540	Good	0.304	Good	24	0.438	Good	-0.071	Revised
5	0.489	Good	0.089	Revised	25	0.436	Good	-0.107	Revised
6	0.446	Good	-0.161	Revised	26	0.385	Good	-0.286	Revised
7	0.438	Good	-0.232	Revised	27	0.383	Good	-0.321	Revised
8	0.453	Good	-0.143	Revised	28	0.416	Good	-0.143	Revised
9	0.414	Good	-0.143	Revised	29	0.458	Good	-0.125	Revised
10	0.409	Good	-0.339	Revised	30	0.385	Good	-0.375	Revised
11	0.438	Good	-0.143	Revised	31	0.404	Good	-0.321	Revised
12	0.436	Good	-0.036	Revised	32	0.433	Good	-0.250	Revised
13	0.400	Good	-0.321	Revised	33	0.441	Good	0,036	Revised
14	0.450	Good	-0.036	Revised	34	0.424	Good	-0.268	Revised
15	0.462	Good	0.250	Revised	35	0.412	Good	-0.321	Revised
16	0.453	Good	-0.089	Revised	36	0.431	Good	-0.304	Revised
17	0.416	Good	-0.143	Revised	37	0.404	Good	-0.232	Revised
18	0.419	Good	-0.196	Revised	38	0.363	Good	-0.482	Revised
19	0.431	Good	-0.232	Revised	39	0.230	Good	-0.929	Revised
20	0.441	Good	-0.089	Revised	40	0.211	Good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

*Analysis of Test Data with Modern Theory**The Unidimensional Assumption Test*

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first-factor Eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first-factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

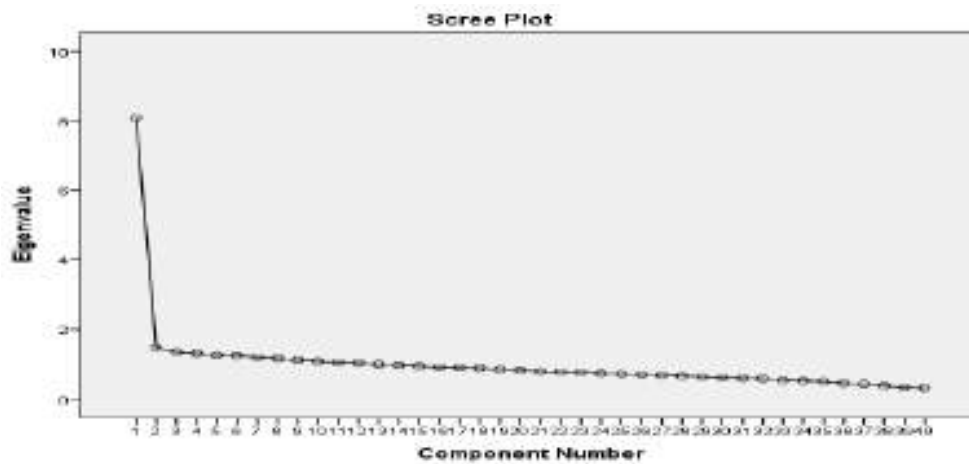


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Correlation is positive, then it can be said that the item fits the model (Sumintono&Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-0.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-0.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-0.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-0.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-0.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-0.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-0.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-0.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-0.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-0.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained were presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono&Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-0.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-0.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-0.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-0.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-0.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-0.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-0.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-0.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-0.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-0.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S.D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

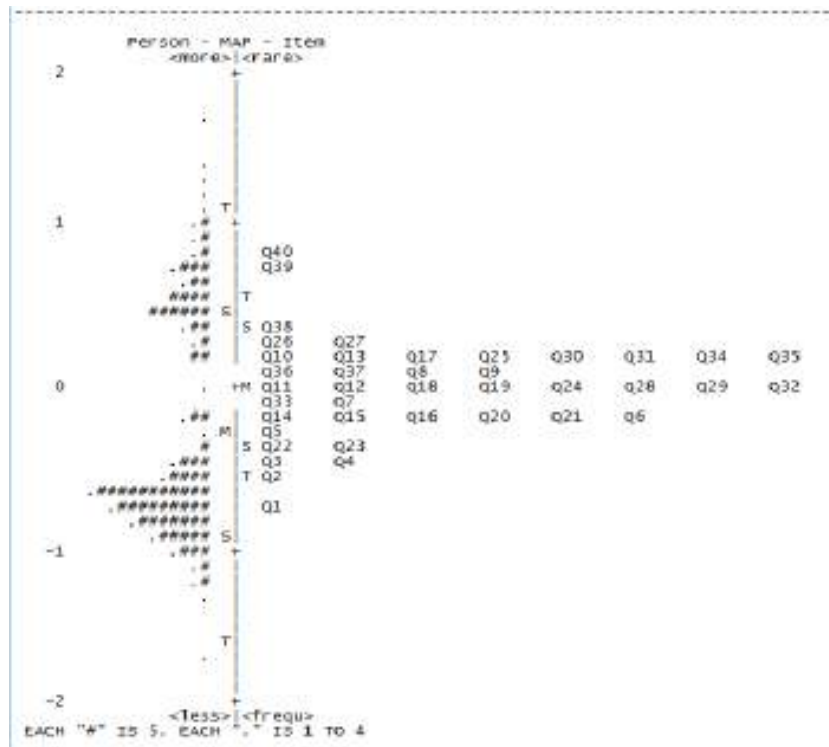


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (PT-Measure Correlation column). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

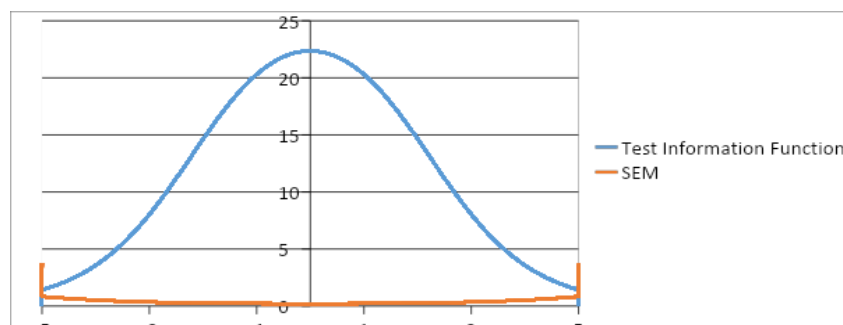


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in the vocational school was based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

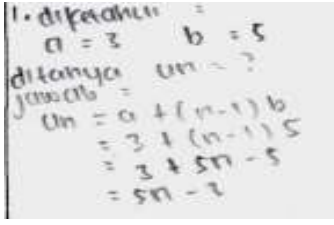
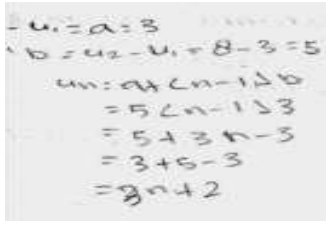
<p>Question 1:</p> <p>Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is ...</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	<p>Pattern 1:</p> 	<p>Pattern 2:</p> 
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term) and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

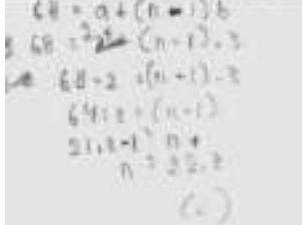
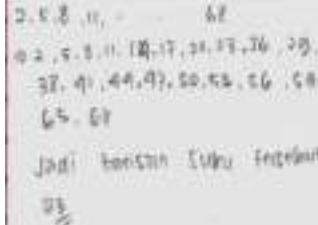
<p>Question 2:</p> <p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>	<p>Pattern 1</p> 	<p>Pattern 2</p> 
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

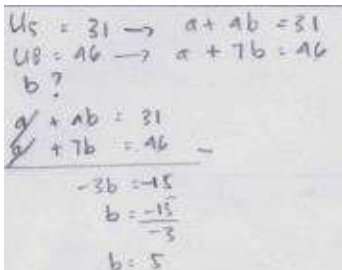
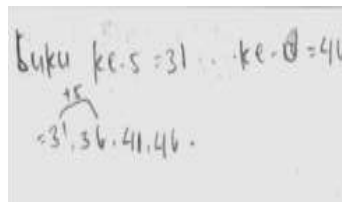
<p>Question 3:</p> <p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>	<p>Pattern 1</p> 	<p>Pattern 2</p> 
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------

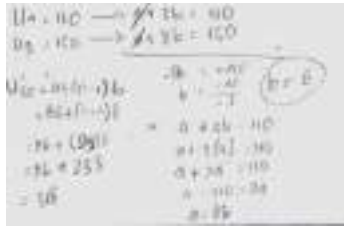
Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

Question 4:
 The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308
 B. 318
 C. 326
 D. 344
 E. 354

Reason:

Pattern 1



Pattern 2

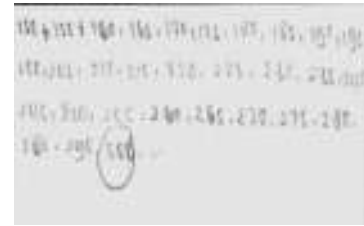


Figure 13. An Example of Student Answers in Item 4

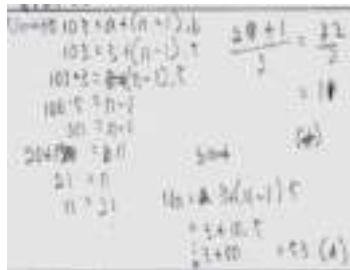
Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

Question 5:

An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
 Then the middle term of the sequence is ...308
 A. 53
 B. 52
 C. 20
 D. 11
 E. 10

Reason:

Answer Pattern 1



Answer Pattern 2



Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice test is easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score. Then, the result is multiplied by 10 to get a value range of 0–10 or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomous makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomous response test is more secondhand and complete than other assessment instruments, such as multiple-choice tests(Gierl et al., 2017) or essay tests(Putri et al., 2020).

Discussion

This research is development research aimed at developing a good mathematical assessment instrument using polytomous responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item

discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of the classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to the modern theory before revising or replacing the item.

Research on learning assessment with open response polytomous was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

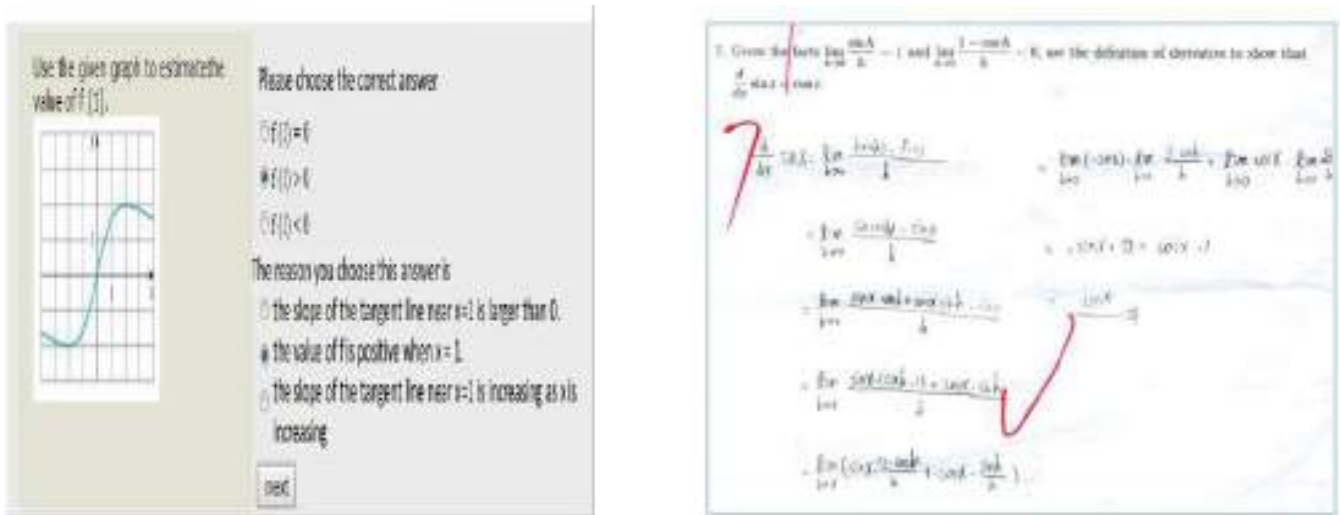


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous

Response Test

Another study on learning assessment with open response polychromous was conducted by Ayanwale(2021).Ayanwale's research compares two polytomous response test analysis methods namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

Other studies related to classical and modern theory were conducted bySarea (2018)and Saepuzaman et al.(2021). Sarea's research states that the response polytomous test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomous test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulasand (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

- A. 308
- B. 318
- C. 326
- D. 344
- E. 354

Reason:

$$\begin{aligned}
 U_4 &= 110 \rightarrow a + 3b = 110 \\
 U_9 &= 150 \rightarrow a + 8b = 150 \\
 \hline
 -5b &= -40 \\
 b &= \frac{-40}{-5} \quad (b = 8) \\
 a + 3(8) &= 110 \\
 a + 24 &= 110 \\
 a &= 110 - 24 \\
 a &= 86 \\
 U_{30} &= a + (n-1)b \\
 &= 86 + (30-1)8 \\
 &= 86 + 233 \\
 &= 318
 \end{aligned}$$

(i)

(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, using a drawing or model, analogy, and formula. Syhlan(2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syhlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomous response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. Thus, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, it is suggested to conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

Authorship Contribution Statement

Sutiarso: Conceptualization, design, development of instruments, analysis, article writing, final approval. Rosidin: Design, development of instruments, analysis, editing. Sulistiawan: Development of instruments, collect data, analysis, editing.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskorandikotomi dan politomidalam teoriresponbutiruntuk pengembangan bank soalmatakuliah matematikadasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://bit.ly/39wv73P>
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campurandikotomus dan politomus generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and polytomous generalized partial credit model]. *Jurnal Suluh Pendidikan*, 17(1), 32-37. <https://bit.ly/3yFQ6eO>
- Arikunto, S. (2012). *Dasar-dasarevaluasipendidikan* [Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://bit.ly/3Nob4mN>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://bit.ly/3lfARKW>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. <https://bit.ly/3MnA9hw>
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geoffrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. ResearchGate. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: Challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>

- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomis dan politomis pada tes prestasi belajar [Equalization of the dichotomous and polytomous mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: What, why, and how? In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <https://bit.ly/3liFMS8>
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <https://bit.ly/3sEcNMA>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://bit.ly/3Psrc8s>
- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://bit.ly/3wljq95>
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://bit.ly/3yIePzb>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standarisasi pendidikan dasar dan menengah* [Standards of content for primary and secondary education]. Indonesian Government publications service. <https://bit.ly/3FZHlhp>
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <https://bit.ly/3MzqkNm>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://bit.ly/37QCtyu>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://bit.ly/3yF4iF5>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya* [Item response theory and its application]. Nuha Medika. <https://bit.ly/39TFDFE>

- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163. <https://doi.org/10.30587/postulat.v1i2.2094>
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisika dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst: Journal of Physics Education and Its Application/Karst: Jurnal Pendidikan Fisika dan Terapannya*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budipekertingkatsekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://bit.ly/3FTtoOU7>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Teori tes klasik and respon [Characteristics of items: Classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <https://doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuluddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data peneliti dengan model rash dan Winstep* [Analysis of research data with rash and Winstep models]. Halaman Moeka Publishing. <https://bit.ly/3NlnGLE>
- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <https://bit.ly/3wjuXp9>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :

Class/Department :

School :

Instructions:Mark (x) one of the correct answer choices, and give the reason

(use another piece of paper to write down your reason)

1. Given an arithmetic sequence: 3, 8, 13, 18, ...The formula for the nth term of the sequence is
- A. $U_n = 5n - 3$ D. $U_n = 4n - 1$
 B. $U_n = 5n - 2$ E. $U_n = 3n + 2$
 C. $U_n = 4n - 1$

Reason:

3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...
- A. 5 D.8
 B. 6 E.11
 C. 7

Reason:

5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...
- A. 53 D.11
 B. 52 E.10
 C. 20

Reason:

7. The nth term of an arithmetic series is $U_n = 3n - 5$. The formula for the sum of the first n terms of the series is ...
- A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
 B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
 C. $S_n = \frac{n}{2}(3n - 4)$

Reason:

9. The sum of all integers between 100 and 300 which are divisible by 5 is ...
- A. 8,200 D.7,600
 B. 8,000 E.7,400
 C. 7,800

Reason:

11. The middle term of an arithmetic sequence

2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...
- A. 12 D.23
 B. 13 E.24
 C. 22

Reason:

4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
- A. 308 D.344
 B. 318 E.354
 C. 326

Reason:

6. Given the arithmetic sequence: 4, 10, 16, 22, ... If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...
- A. 18 D.24
 B. 20 E.26
 C. 22

Reason:

8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....
- A. $5n - 20$ D. $2n - 20$
 B. $5n - 10$ E. $2n - 10$
 C. $2n - 30$

Reason:

10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?
- A. 24 D.27
 B. 25 E.28
 C. 26

Reason:

12. A number of candies are distributed among

is 25.If the difference is 4 and the 5th term is 21.Then the sum of all the terms in the sequence is ...

- A. 175D.295
- B. 189E.375
- C. 275

Reason:

five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces

- A. 60D.75
- B. 65E.80
- C. 70

Reason:

13. The sum of the first n terms of a series is $2n^2 - n$. So the 12th term of the series is...

- A. 564D.45
- B. 276E.36
- C. 48

Reason:

14. The number of terms in the geometric sequence: 3, 6, 12, ..., 3072 is ...

- A. 9D.12
- B. 10E.13
- C. 11

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...

- A. 32D.256
- B. 64E.512
- C. 128

Reason:

16. The value of the middle term of the geometric sequence:

- 6, 3, ..., $\frac{3}{512}$ is ...
- A. $\frac{1}{16}$ D. $\frac{4}{16}$
 - B. $\frac{2}{16}$ E. $\frac{5}{16}$
 - C. $\frac{3}{16}$

Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm

- A. 18D.35
- B. 24E.40.5
- C. 27.5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...

- A. $\frac{3}{4}$ D. $-\frac{1}{2}$
- B. $\frac{1}{4}$ E. $-\frac{3}{4}$
- C. $\frac{1}{3}$

Reason:

19. A ball falls from a height of 10 m and bounces back $\frac{3}{4}$ times its previous height. The total number of paths until the ball stops is.... m

- A. 60D. 90
- B. 70E. 100
- C. 80

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b + a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.

- The value of $3a + b$ is ...
- A. 8D. 14
 - B. 10E. 20
 - C. 12

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$. If $K = L$, then cis ...

- A. 12D. 15
- B. 13E. 16
- C. 14

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$. Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$
- D. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ is...

- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ is...

- A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$

- is ...
 A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then

- matrix determinant A is ...
 A. 0 D. 2
 B. 1 E. 4
 C. 2

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$.

Inverse matrix $(AB)^{-1}$ is ...

- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$

Reason:

33. The roots of the quadratic equation

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$,
 and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.

If $A + B = C$, then $x + y = \dots$

- A. -5 D. 3
 B. -1 E. 5
 C. 1

Reason:

26. If $A = \begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and

$C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$ then $A(B - C) = \dots$

- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix.

Value of x that satisfies is ...

- A. -5 D. 3
 B. -4 E. 4
 C. -3

Reason:

30. Transpose matrix P is P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then

matrix $(P^T)^{-1}$ is ...

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$

Reason:

32. The roots of the quadratic equation $3x^2 - 4x + 3 = 0$ are ...

- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$

Reason:

34. The roots of the equation $x^2 - 2x - 3 = 0$

$5x^2 + 4x - 12 = 0$ are ...

- A. -2 and $-\frac{5}{6}$
- B. 2 and $-\frac{5}{6}$
- C. 2 and $\frac{6}{5}$
- D. -2 and $-\frac{6}{5}$
- E. -2 and $\frac{6}{5}$

Reason:

35. The roots of the quadratic equation: $2x^2 - 3x - 9 = 0$ are x_1 and x_2 . Value of $x_1^2 + x_2^2$ is ...

- A. $11\frac{1}{4}$
- B. $6\frac{3}{4}$
- C. $2\frac{1}{4}$
- D. $-6\frac{3}{4}$
- E. $-11\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...

- A. $y = x^2 - 2x + 1$
- B. $y = x^2 - 2x + 3$
- C. $y = x^2 + 2x - 1$
- D. $y = x^2 + 2x + 1$
- E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0,-4)$ then the value of $f(7)$ is ...

- A. -16
- B. -17
- C. -18
- D. -19
- E. -20

Reason:

are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...

- A. -4
- B. -2
- C. 0
- D. 2
- E. 4

Reason:

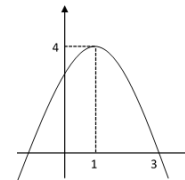
36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α and β . The quadratic equation $(\alpha - 2)$ and $(\beta - 2)$ is ...

- A. $x^2 + 6x + 5 = 0$
- B. $x^2 + 6x + 7 = 0$
- C. $x^2 + 6x + 11 = 0$
- D. $x^2 - 2x + 3 = 0$
- E. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation? ...

- A. $y = x^2 + 2x + 3$
- B. $y = x^2 - 2x - 3$
- C. $y = -x^2 + 2x - 3$
- D. $y = -x^2 - 2x + 3$
- E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...

- A. $y = -x^2 + 2x - 3$
- B. $y = -x^2 + 2x + 3$
- C. $y = -x^2 - 2x + 3$
- D. $y = -x^2 - 2x - 5$
- E. $y = -x^2 - 2x + 5$

Reason:



Ok_EU-JER_11_3_1441_SUTIARSO_FINAL2.docx
1804K



European Journal of Educational Research

Volume 11, Issue 3, 1441 - 1462.

ISSN: 2165-8714

<https://www.eu-jer.com/>

Developing Assessment Instrument Using Polytomous Response in Mathematics

SugengSutiarso*

University of Lampung, INDONESIA

UndangRosidin

University of Lampung, INDONESIA

AanSulistiawan

Vocational School, INDONESIA

Received: November 25, 2021 • Revised: February 2, 2022 • Accepted: May 17, 2022

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used an open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) the open polytomous response test has a good category according to classical and modern theory. However, the discrimination power of test items in classical theory needs several revisions, (2) the assessment instrument using the polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses.

To cite this article: Sutiarso, S., Rosidin, U., & Sulistiawan, A. (2022). Developing assessment instrument using polytomous response in mathematics. *European Journal of Educational Research*, 11(3), 1441-1462. <https://doi.org/10.12973/eu-jer.11.3.1441>

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools was divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).



Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with specific rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over essay test are firstly, the test can be

* Corresponding author:

SugengSutiarso, University of Lampung, Sumantri Brojonegoro, Bandar Lampung, Indonesia. ✉ sugeng.sutiarso@fkip.unila.ac.id

© 2022 The Author(s). **Open Access**- This article is under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>)



conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weaknesses. The multiple-choice test is not able to portray the actual abilities of students, the answers of this test tend to be guessing games or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). A multiple-choice test with only two answer choices is called a "dichotomous test," and a multiple-choice test with more than two answer choices is called a "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or henceforth called as polytomous response test (Suwarto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as a test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), a test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into an open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response tests for students in college and senior or junior school. Students in college and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools are not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomous test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmalawati, 2020). Therefore, the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomous response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing the item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because, in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows: (1) Does

the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory? and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's (2013) model, with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

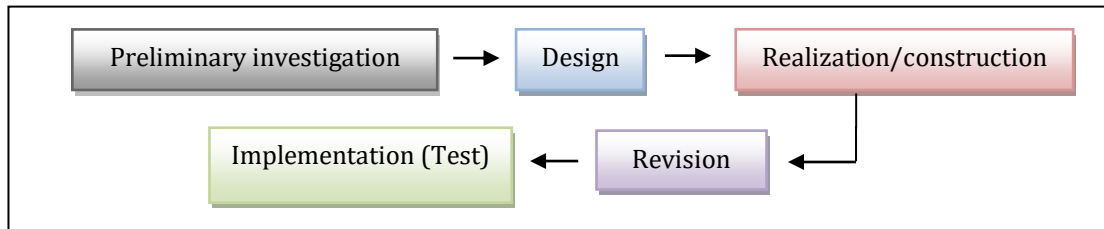


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the

indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. Questionnaire data analysis (qualitative analysis)

There are two sets of questionnaire data namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically, for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. Test data analysis (empirical analysis)

After conducting the content validity test, the researcher conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to have good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program was used for classical theory

and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1. Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2. Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep) because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Correlation column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit's MNSQ value is 0.5 to 1.5 and the Pt-Measure Correlation is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay tests and 20% of teachers used multiple-choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

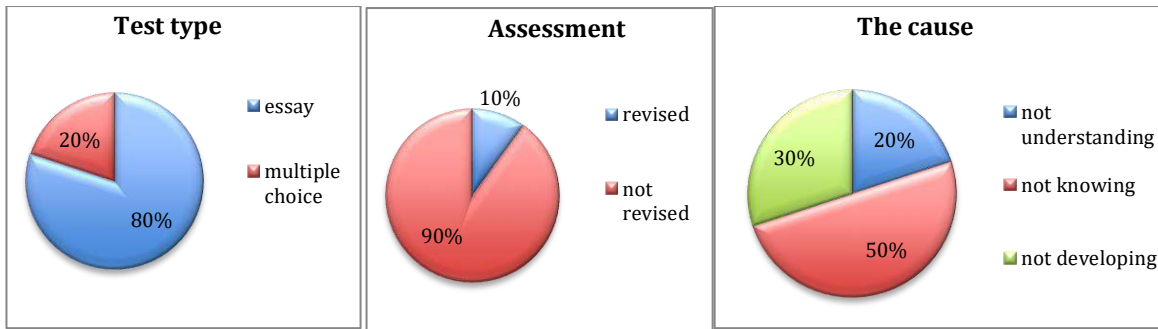


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	Good	0.196	Revised	21	0.482	Good	0.143	Revised
2	0.487	Good	0.179	Revised	22	0.535	Good	0.429	Good
3	0.528	Good	0.214	Revised	23	0.492	Good	0.250	Revised
4	0.540	Good	0.304	Good	24	0.438	Good	-0.071	Revised
5	0.489	Good	0.089	Revised	25	0.436	Good	-0.107	Revised
6	0.446	Good	-0.161	Revised	26	0.385	Good	-0.286	Revised
7	0.438	Good	-0.232	Revised	27	0.383	Good	-0.321	Revised
8	0.453	Good	-0.143	Revised	28	0.416	Good	-0.143	Revised
9	0.414	Good	-0.143	Revised	29	0.458	Good	-0.125	Revised
10	0.409	Good	-0.339	Revised	30	0.385	Good	-0.375	Revised
11	0.438	Good	-0.143	Revised	31	0.404	Good	-0.321	Revised
12	0.436	Good	-0.036	Revised	32	0.433	Good	-0.250	Revised
13	0.400	Good	-0.321	Revised	33	0.441	Good	0,036	Revised
14	0.450	Good	-0.036	Revised	34	0.424	Good	-0.268	Revised
15	0.462	Good	0.250	Revised	35	0.412	Good	-0.321	Revised
16	0.453	Good	-0.089	Revised	36	0.431	Good	-0.304	Revised
17	0.416	Good	-0.143	Revised	37	0.404	Good	-0.232	Revised
18	0.419	Good	-0.196	Revised	38	0.363	Good	-0.482	Revised
19	0.431	Good	-0.232	Revised	39	0.230	Good	-0.929	Revised
20	0.441	Good	-0.089	Revised	40	0.211	Good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

*Analysis of Test Data with Modern Theory**The Unidimensional Assumption Test*

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first-factor Eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first-factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

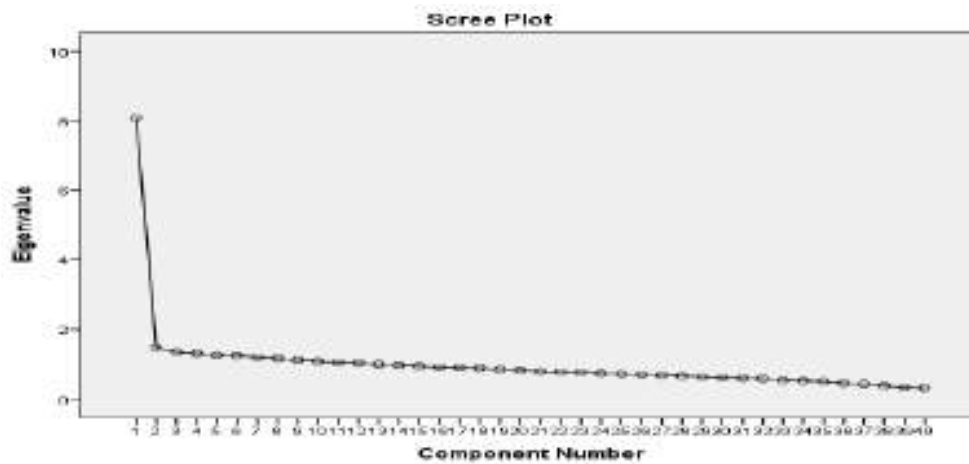


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Correlation is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-0.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-0.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-0.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-0.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-0.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-0.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-0.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-0.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-0.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-0.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained were presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-0.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-0.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-0.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-0.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-0.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-0.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-0.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-0.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-0.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-0.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-0.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-0.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

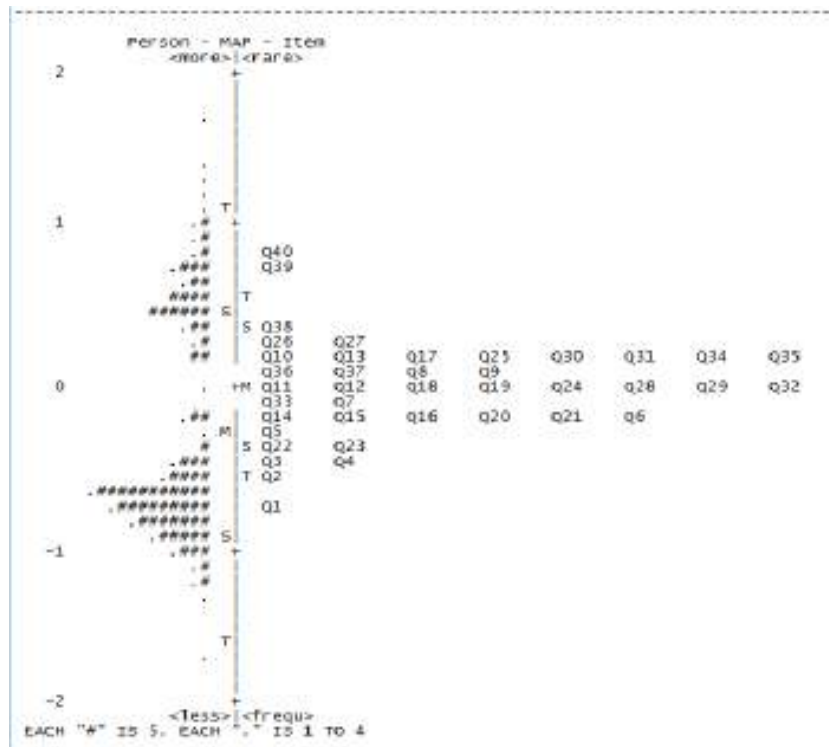


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (PT-Measure Correlation column). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

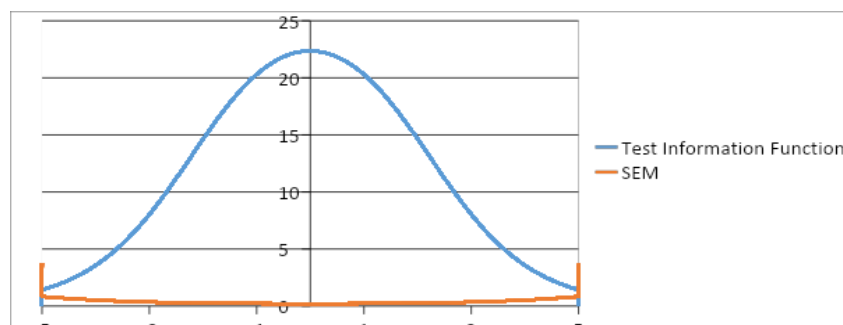


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in the vocational school was based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

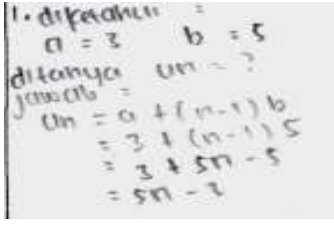
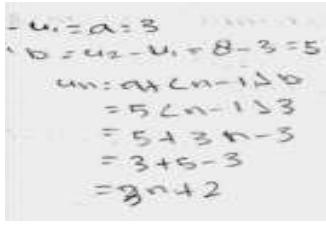
<p>Question 1:</p> <p>Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is ...</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	<p>Pattern 1:</p> 	<p>Pattern 2:</p> 
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term) and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

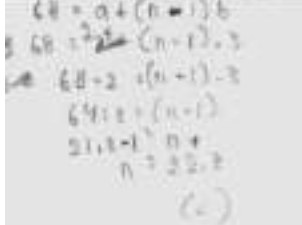
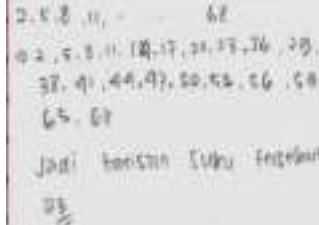
<p>Question 2:</p> <p>Given an arithmetic sequence: 2, 5, 8, 11,, 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>	<p>Pattern 1</p> 	<p>Pattern 2</p> 
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

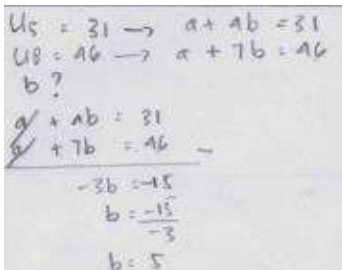
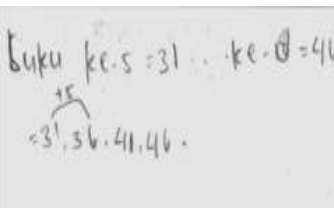
<p>Question 3:</p> <p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>	<p>Pattern 1</p> 	<p>Pattern 2</p> 
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------

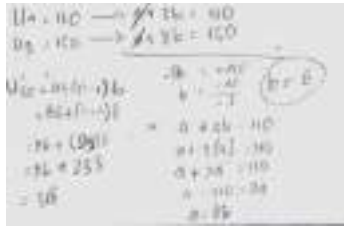
Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

Question 4:
 The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308
 B. 318
 C. 326
 D. 344
 E. 354

Reason:

Pattern 1



Pattern 2

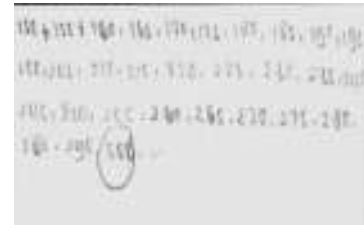


Figure 13. An Example of Student Answers in Item 4

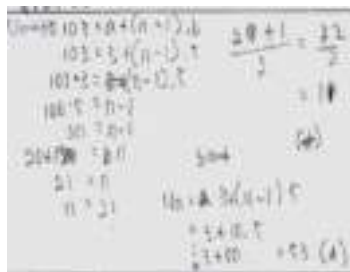
Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

Question 5:

An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
 Then the middle term of the sequence is ...308
 A. 53
 B. 52
 C. 20
 D. 11
 E. 10

Reason:

Answer Pattern 1



Answer Pattern 2



Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice test is easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score. Then, the result is multiplied by 10 to get a value range of 0–10 or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomous makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomous response test is more secondhand and complete than other assessment instruments, such as multiple-choice tests(Gierl et al., 2017) or essay tests(Putri et al., 2020).

Discussion

This research is development research aimed at developing a good mathematical assessment instrument using polytomous responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item

discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of the classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to the modern theory before revising or replacing the item.

Research on learning assessment with open response polytomous was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

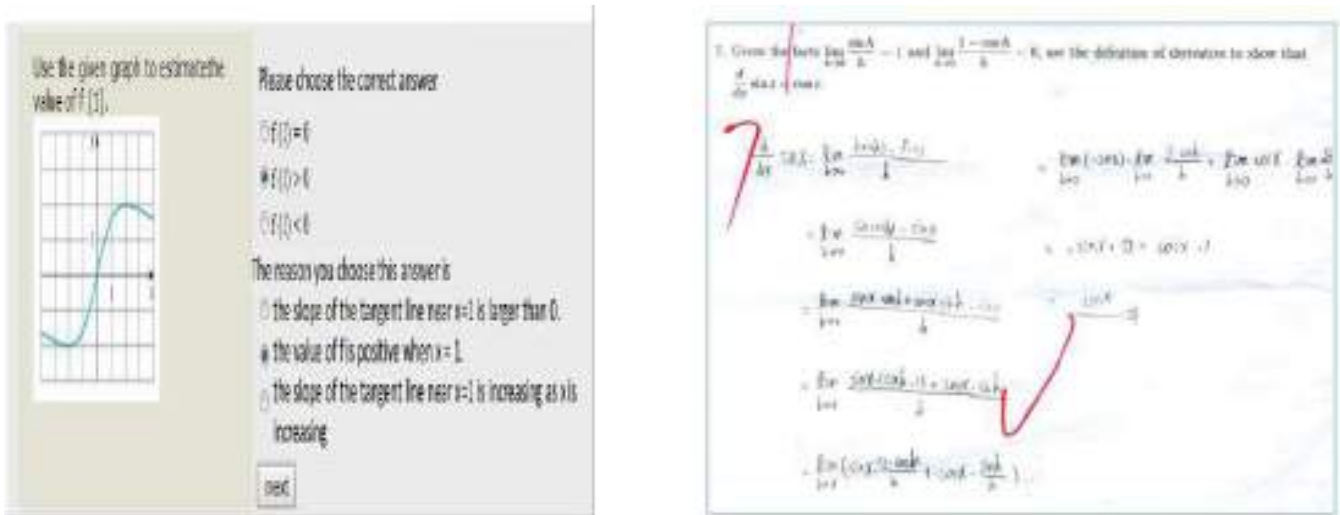


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous

Response Test

Another study on learning assessment with open response polychromous was conducted by Ayanwale(2021).Ayanwale's research compares two polytomous response test analysis methods namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

Other studies related to classical and modern theory were conducted bySarea (2018)and Saepuzaman et al.(2021). Sarea's research states that the response polytomous test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomous test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

- A. 308
- B. 318
- C. 326
- D. 344
- E. 354

Reason:

$$\begin{aligned}
 U_4 &= 110 \rightarrow a + 3b = 110 \\
 U_9 &= 150 \rightarrow a + 8b = 150 \\
 \hline
 & -5b = -40 \\
 & b = \frac{-40}{-5} \quad (b = 8) \\
 & a + 3(8) = 110 \\
 & a + 24 = 110 \\
 & a = 110 - 24 \\
 & a = 86 \\
 U_{30} &= a + (n-1)b \\
 &= 86 + (29)8 \\
 &= 86 + 233 \\
 &= 318
 \end{aligned}$$

(i)

$$\begin{aligned}
 &150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200 \\
 &205, 210, 215, 220, 225, 230, 235, 240, 245 \\
 &250, 255, 260, 265, 270, 275, 280, \\
 &285, 290, 295, 300
 \end{aligned}$$

(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, using a drawing or model, analogy, and formula. Syhlan(2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syhlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomous response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. Thus, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, it is suggested to conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

Authorship Contribution Statement

Sutiarso: Conceptualization, design, development of instruments, analysis, article writing, final approval. Rosidin: Design, development of instruments, analysis, editing. Sulistiawan: Development of instruments, collect data, analysis, editing.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://bit.ly/39wv73P>
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomis dan politomis generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and polytomous generalized partial credit model]. *JurnalSuluh Pendidikan*, 17(1), 32-37. <https://bit.ly/3yFQ6eO>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://bit.ly/3Nob4mN>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://bit.ly/3lfARkW>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. <https://bit.ly/3MnA9hw>
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geoffrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. ResearchGate. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: Challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>

- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar [Equalization of the dichotomous and polytomous mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: What, why, and how? In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <https://bit.ly/3liFMS8>
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <https://bit.ly/3sEcNMA>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://bit.ly/3Psrc8s>
- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://bit.ly/3wljq95>
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://bit.ly/3yIePzb>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Standards of content for primary and secondary education]. Indonesian Government publicationservice. <https://bit.ly/3FZHlhp>
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <https://bit.ly/3MzqkNm>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://bit.ly/37QCtyu>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://bit.ly/3yF4iF5>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya* [Item response theory and its application]. NuhaMedika. <https://bit.ly/39TFDIF>

- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163. <https://doi.org/10.30587/postulat.v1i2.2094>
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst: Journal of Physics Education and Its Application/Karst: Jurnal Pendidikan Fisika dan Terapannya*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://bit.ly/3FTouU7>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Teori tes klasik and respon [Characteristics of items: Classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <https://doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan Winstep* [Analysis of research data with rash and Winstep models]. Halaman Moeka Publishing. <https://bit.ly/3NlnGLc>
- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <https://bit.ly/3wjuXp9>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :

Class/Department :

School :

Instructions:Mark (x) one of the correct answer choices, and give the reason

(use another piece of paper to write down your reason)

1. Given an arithmetic sequence: 3, 8, 13, 18, ...The formula for the nth term of the sequence is

A. $U_n = 5n - 3$ D. $U_n = 4n - 1$
 B. $U_n = 5n - 2$ E. $U_n = 3n + 2$
 C. $U_n = 4n - 1$

Reason:

3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...

A. 5 D. 8
 B. 6 E. 11
 C. 7

Reason:

5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103. Then the middle term of the sequence is ...

A. 53 D. 11
 B. 52 E. 10
 C. 20

Reason:

7. The nth term of an arithmetic series is $U_n = 3n - 5$. The formula for the sum of the first n terms of the series is ...

A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
 B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
 C. $S_n = \frac{n}{2}(3n - 4)$

Reason:

9. The sum of all integers between 100 and 300 which are divisible by 5 is ...

A. 8,200 D. 7,600
 B. 8,000 E. 7,400
 C. 7,800

Reason:

11. The middle term of an arithmetic sequence

2. Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...

A. 12 D. 23
 B. 13 E. 24
 C. 22

Reason:

4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

A. 308 D. 344
 B. 318 E. 354
 C. 326

Reason:

6. Given the arithmetic sequence: 4, 10, 16, 22, ... If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...

A. 18 D. 24
 B. 20 E. 26
 C. 22

Reason:

8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....

A. $5n - 20$ D. $2n - 20$
 B. $5n - 10$ E. $2n - 10$
 C. $2n - 30$

Reason:

10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?

A. 24 D. 27
 B. 25 E. 28
 C. 26

Reason:

12. A number of candies are distributed among

is 25.If the difference is 4 and the 5th term is 21.Then the sum of all the terms in the sequence is ...

- A. 175D.295
- B. 189E.375
- C. 275

Reason:

five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces

- A. 60D.75
- B. 65E.80
- C. 70

Reason:

13. The sum of the first n terms of a series is $2n^2-n$.So the 12th term of the series is...

- A. 564D.45
- B. 276E.36
- C. 48

Reason:

14. The number of terms in thegeometric sequence:3, 6, 12, ..., 3072 is ...

- A. 9D.12
- B. 10E.13
- C. 11

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...

- A. 32D.256
- B. 64E.512
- C. 128

Reason:

16. The value of the middle term of the geometric sequence:

6, 3, ..., $3/512$ is ...

- A. $\frac{1}{16}$ D. $\frac{4}{16}$
- B. $\frac{2}{16}$ E. $\frac{5}{16}$
- C. $\frac{3}{16}$

Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm

- A. 18D.35
- B. 24E.40.5
- C. 27.5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...

- A. $\frac{3}{4}$ D. $-\frac{1}{2}$
- B. $\frac{1}{4}$ E. $-\frac{3}{4}$
- C. $\frac{1}{3}$

Reason:

19. A ball falls from a height of 10 m and bounces back $3/4$ times its previous height. The total number of paths until the ball stops is.... m

- A. 60D. 90
- B. 70E. 100
- C. 80

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.

The value of $3a + b$ is ...

- A. 8D. 14
- B. 10E. 20
- C. 12

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$. If $K = L$, then cis ...

- A. 12D. 15
- B. 13E. 16
- C. 14

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$. Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$
- C. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ is...

- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ is...

- A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...

- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ...

- A. 0 D. 2
 B. 1 E. 4
 C. 2

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$. Inverse matrix $(AB)^{-1}$ is ...

- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$

Reason:

33. The roots of the quadratic equation

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.

If $A + B = C$, then $x + y = \dots$

- A. -5 D. 3
 B. -1 E. 5
 C. 1

Reason:

26. If $A = \begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$ then $A(B - C) = \dots$

- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix. Value of x that satisfies is ...

- A. -5 D. 3
 B. -4 E. 4
 C. -3

Reason:

30. Transpose matrix P is P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ...

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$

Reason:

32. The roots of the quadratic equation $3x^2 - 4x + 3 = 0$ are ...

- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$

Reason:

34. The roots of the equation $x^2 - 2x - 3 = 0$

$5x^2 + 4x - 12 = 0$ are ...

- A. -2 and $\frac{5}{6}$
- B. 2 and $-\frac{5}{6}$
- C. 2 and $\frac{6}{5}$
- D. -2 and $-\frac{6}{5}$
- E. -2 and $\frac{6}{5}$

Reason:

35. The roots of the quadratic equation: $2x^2 - 3x - 9 = 0$ are x_1 and x_2 . Value of $x_1^2 + x_2^2$ is ...

- A. $11\frac{1}{4}$
- B. $6\frac{3}{4}$
- C. $2\frac{1}{4}$
- D. $-6\frac{3}{4}$
- E. $-11\frac{1}{4}$

Reason:

37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...

- A. $y = x^2 - 2x + 1$
- B. $y = x^2 - 2x + 3$
- C. $y = x^2 + 2x - 1$
- D. $y = x^2 + 2x + 1$
- E. $y = x^2 + 2x + 3$

Reason:

39. If f is a quadratic function whose graph passes through the points $(1,0)$, $(4,0)$ and $(0,-4)$ then the value of $f(7)$ is ...

- A. -16
- B. -17
- C. -18
- D. -19
- E. -20

Reason:

are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...

- A. -4
- B. -2
- C. 0
- D. 2
- E. 4

Reason:

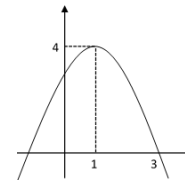
36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α and β . The quadratic equation $(\alpha - 2)$ and $(\beta - 2)$ is ...

- A. $x^2 + 6x + 5 = 0$
- B. $x^2 + 6x + 7 = 0$
- C. $x^2 + 6x + 11 = 0$
- D. $x^2 - 2x + 3 = 0$
- E. $x^2 + 2x + 11 = 0$

Reason:

38. The figure below is a graph of the quadratic equation? ...

- A. $y = x^2 + 2x + 3$
- B. $y = x^2 - 2x - 3$
- C. $y = -x^2 + 2x - 3$
- D. $y = -x^2 - 2x + 3$
- E. $y = -x^2 + 2x + 3$



Reason:

40. The graph equation of a quadratic function has an extreme point $(-1, 4)$ and through $(0, 3)$ is ...

- A. $y = -x^2 + 2x - 3$
- B. $y = -x^2 + 2x + 3$
- C. $y = -x^2 - 2x + 3$
- D. $y = -x^2 - 2x - 5$
- E. $y = -x^2 - 2x + 5$

Reason:



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

3rd Final Paper (ID#21112502244011)

2 messages

Editor - European Journal of Educational Research <editor@eu-jer.com>
To: SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Sun, May 22, 2022 at 1:19 PM

Dear Dr. Sutiarso,

Thank you for your email. We have updated your paper.

Please find the attached 3rd finalized paper will be published.

We will publish it as online first soon.

The official publication date of your paper is July 15, 2022.

Best regards,
Ahmet Savas Ph.D.
Editor- European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 5/22/2022 6:04 AM, SUGENG SUTJARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research.

I have checked that there are no more errors in the article substantially. However, I found a few writing errors in the answer choices (in the appendix), namely there is no space between the answer choices (as shown below).

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :

Class/Department :

School :

Instructions: Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

1. Given an arithmetic sequence: 3, 8, 13, 18, ...
The formula for the n th term of the sequence is ...
A. $U_n = 5n - 3$ D. $U_n = 4n - 1$
B. $U_n = 5n - 2$ E. $U_n = 3n + 2$
C. $U_n = 4n - 1$
- Reason:

2. Given an arithmetic sequence: 2, 5, 8, 11, ...
68. The number of terms in the sequence is...
A. 12 D. 23
B. 13 E. 24
C. 22
- Reason:

3. An arithmetic sequence, the 5th term is 31
4. The 4th and 9th terms of an arithmetic



Appendix

Instrument of the Open Polytomous Response Test

Name of Student :

Class/Department :

School :

Instructions: Mark (x) one of the correct answer choices, and give the reason
(use another piece of paper to write down your reason)

1. Given an arithmetic sequence: 3, 8, 13, 18, ...
The formula for the n th term of the sequence is ...
A. $U_n = 5n - 3$ D. $U_n = 4n - 1$
B. $U_n = 5n - 2$ E. $U_n = 3n + 2$
C. $U_n = 4n - 1$
- Reason:

2. Given an arithmetic sequence: 2, 5, 8, 11, ...
68. The number of terms in the sequence is...
A. 12 D. 23
B. 13 E. 24
C. 22
- Reason:



I revised it, and here I resubmit the corrected article.
Hopefully, this article is final (no more mistakes).

Thank you for your kindness and patience in revising my article.

Best regards,
Sugeng Sutiarmo
University of Lampung

On Sat, May 21, 2022 at 7:58 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarmo,

Thank you for your email. We have updated your paper.

Please find the attached 2nd finalized paper will be published.

We will publish it as online first soon.

The official publication date of your paper is July 15, 2022.

Best regards,
Ahmet Savas Ph.D.
Editor- European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com

On 5/21/2022 1:09 PM, SUGENG SUTIARSO wrote:

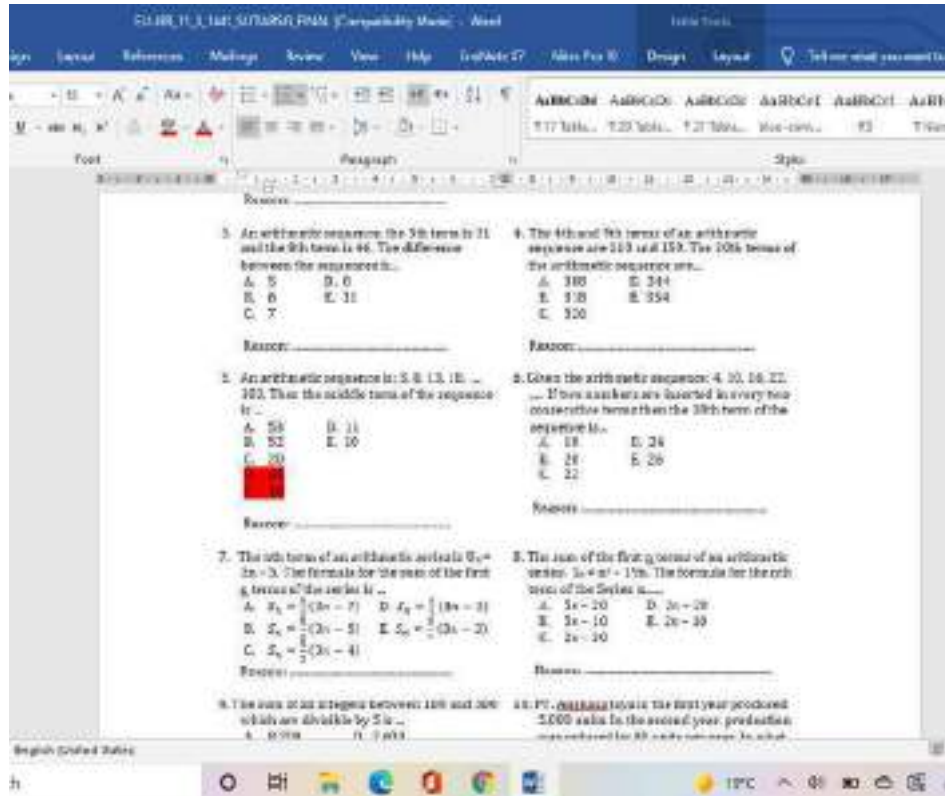
Dear Ahmet C. Savas, Ph.D.

Editor, European Journal of Educational Research

I have checked again, there is still 1 typing error, namely: in appendix no 5 there is double writing on the answer choices D and E (highlighted in red), and I have deleted the answer choices.

Here I resubmit the article that I have revised (ok_final article).

Thank you for your kindness in revising my article.



Best regards,

Sugeng Sutiarmo

University of Lampung

On Sat, May 21, 2022 at 3:45 AM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarmo,

Thank you for your email. We have updated your paper. We have corrected a few mistakes.

Please find the attached finalized paper will be published.

We will publish it as online first soon.

The official publication date of your paper is July 15, 2022.

Best regards,

Ahmet Savas Ph.D.

Editor- European Journal of Educational Research

editor@eu-jer.com

www.eu-jer.com

On 5/19/2022 5:14 PM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

I checked the language and references of my paper, and there are some edited parts of the appendix (highlighted in green), namely: (1) the word "and" (previously, the word "dan" in Indonesian) and (2) the word "is" (previously, the word "adalah" in Indonesian).
Here is my final paper attached.

Best regards,
Sugeng Sutiarto
Lampung University

On Tue, May 17, 2022 at 6:12 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarto,

Please see the attached galley proof of your paper (ID#2103030737) (word file). Please highlight in green for your edited parts.

By the way,

- 1- Please check the language of your paper as a proofreading lastly.
- 2- Please check all references regarding with attached citation guide for APA 7 style. (Please see the citation guide page in our web site: <https://www.eujem.com/citation-guide>)

We ask you to check it please. Please edit at word file and resend it to me please in 2 days.

We are looking forward to getting your final paper by **May 19, 2022**.

Best regards,
Ahmet Savas Ph.D.
Editor, European Journal of Educational Management

<http://www.eujem.com>

editor@eujem.com

On 5/10/2022 1:24 PM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

Here I attach a revision of your proofreading suggestion.

Best regards,
Sugeng Sutiarto
Lampung University

On Fri, May 6, 2022 at 7:39 PM Editor - European Journal of Educational Research <editor@eu-jer.com> wrote:

Dear Dr. Sutiarto,

Thank you for your kind email. Please see the attached file as the proofreading of your paper.

We are preparing the galley proof of your paper.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

editor@eu-jer.com
www.eu-jer.com

On 4/23/2022 4:36 AM, SUGENG SUTIARSO wrote:

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of
Educational Research

Here, I attach the copyright transfer
agreement.

Best regards,
Sugeng Sutiarto
Lampung University

On Fri, Apr 22, 2022 at 5:24 PM
European Journal of Educational
Research <editor@eu-jer.com> wrote:

Dear Dr. Sugeng Sutiarto,

We have received your payment
about your paper entitled
"Development of Mathematics
Assessment Instruments for
Learning with Polytomous Response
in Vocational School"
ID#21112502244011. Thanks.

We kindly ask from you to sign the
copyright transfer agreement for
your paper. After all author(s)
signed, please scan and send via
email to me **as soon as possible**.
Please download the pdf file of this
agreement from this link : <https://eu-jer.com/EU-JER-copyright-transfer-agreement.pdf> You can use e-
signature, if you have. Also you can
use your mobil phone as a scanner.
If the other author live in another
city, he/she sign the paper and send
this paper via email. Than you can
sign on this paper.

We are preparing the galley proof of
your paper. We will send it to you in
order to check before publication.
The preparing of galley proofs may
take some time because of our
intensity. Thank you for your
patience.

We are looking forward to getting
copyright transfer agreement.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of
Educational Research
editor@eu-jer.com
www.eu-jer.com



EU-JER_11_3_1441_SUTIARSO_FINAL3.docx

1804K



European Journal of Educational Research

Volume 11, Issue 3, 1441 - 1462.

ISSN: 2165-8714

<https://www.eu-jer.com/>

Developing Assessment Instrument Using Polytomous Response in Mathematics

SugengSutiarso*

University of Lampung, INDONESIA

UndangRosidin

University of Lampung, INDONESIA

AanSulistiawan

Vocational School, INDONESIA

Received: November25, 2021 • Revised: February2, 2022 • Accepted: May 17, 2022

Abstract: This research is a developmental research aiming at developing a good mathematical test instrument using polytomous responses based on classical and modern theories. This research design uses the Plomp model, which consists of five stages, (1) preliminary investigation, (2) design, (3) realization/construction, (4) revision, and (5) implementation (testing). The study was conducted in three vocational schools in Lampung Province, Indonesia. The study involved 413 students, consisting of 191 male and 222 female students. The data were collected through questionnaire and test. The questionnaire was used to identify the assessment instruments currently employed by teachers and to be validated by the experts of mathematics and educational evaluation. The test used an open polytomous response test numbering of 40 items. The data were analyzed using both classical and modern theories. The results show that (1) the open polytomous response test has a good category according to classical and modern theory. However, the discrimination power of test items in classical theory needs several revisions, (2) the assessment instrument using the polytomous response of open multiple choice can guarantee information on the actual competence of students. This is proven by the fact that there is a harmony between the analysis result obtained from classical and modern theory from the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Keywords: Assessment instrument, classical and modern theory, vocational school, polytomous responses.

To cite this article: Sutiarso, S., Rosidin, U., & Sulistiawan, A. (2022). Developing assessment instrument using polytomous response in mathematics. *European Journal of Educational Research*, 11(3), 1441-1462. <https://doi.org/10.12973/eu-jer.11.3.1441>

Introduction

Assessment is an important activity that needs to be administered by teachers in schools. The conventional paradigm often interprets assessment as a way to find out the achievement of student learning outcomes as a whole, so that the assessment is positioned as a separate activity from the learning process (Syaifuddin, 2020). Referring to the current paradigm, assessment in schools was divided into three types, e.g., assessment as learning, assessment for learning, and assessment of learning (Wulan, 2018). The three types of assessments aim to provide recognition of the achievement of student learning outcomes after the learning process (Earl, 2013). Below is the assessment pyramid (Figure 1).

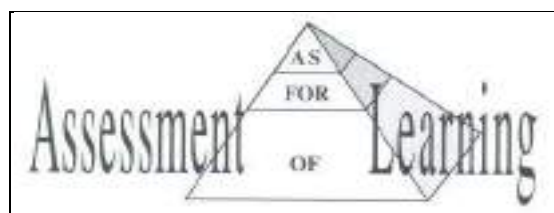


Figure 1. Assessment Pyramid

Assessment can be administered through a test. A test is a tool or procedure used to find out or measure students' abilities in particular areas with specific rules (Arikunto, 2012). A test consists of two types, namely multiple-choice and essay. A multiple-choice test is a form of assessment in which each item provides options, and one of the options is the correct answer. An essay test is a form of assessment that requires answers in sentences or words. Each type of test has its own strengths or weaknesses. The strengths of multiple-choice test over essay test are firstly, the test can be

*Corresponding author:

SugengSutiarso, University of Lampung, Sumantri Brojonegoro, Bandar Lampung, Indonesia. ✉ sugeng.sutiarso@fkip.unila.ac.id



conducted for many students, it is more objective, and the test results can be obtained more quickly. In addition to the strengths, it has several weaknesses. The multiple-choice test is not able to portray the actual abilities of students, the answers of this test tend to be guessing games or trial and error (Rosidin, 2017). In addition, the strength of the multiple-choice test has a scoring certainty compared to the essay test, namely 1 and 0 (score 1 for the correct answer, and 0 for the wrong answer). A multiple-choice test with only two answer choices is called a "dichotomous test," and a multiple-choice test with more than two answer choices is called a "polytomous test" (Kartono, 2008).

Until quite recently, multiple-choice test has been widely used by teachers to assess students' abilities, especially those with a large number and a wide area of expertise. To reduce the weakness of multiple-choice test, in the last four decades, experts have developed multiple-choice test by combining multiple-choice test and essay into multiple-choice test with reasons or henceforth called as polytomous response test (Suwarto, 2012). The polytomous response test score ranges from 1 – 4 in which score 4 for the correct answer and reason, score 3 for the correct answer but the wrong reason, score 2 for the wrong answer but the correct reason, and score 1 for the wrong answer and reason (Kartono, 2008).

In the 1980s, the first test that was focused on and developed by experts was the closed polytomous response test, also known as the two-tier test (Treagust, 1988). This test consists of two levels: the first is choosing answers on the multiple-choice test, and the second level is choosing reasons based on the answer choices at the first level (Chandrasegaran et al., 2007). Several studies on the closed polytomous response test have been carried out, such as a test on mathematical ability in middle school (Hilton et al., 2013; Rovita et al., 2020), a test on calculus material (Khiyarunnisa & Retnawati, 2018), test on higher order thinking mathematical skills (Sundari, et al., 2021), and test on mathematical connection material (Lestari et al., 2021). Although the closed polytomous response test has been widely developed, researchers have found weaknesses in the test, such as students' misconceptions or students' actual competence cannot be identified in detail (Antara et al., 2019), the test instrument is difficult to construct (Khusnah, 2019), and student answers are still guessed (Myanda et al., 2020). However, there are strengths in the closed polytomous response test, such as the consistency of student answers errors, which is easily observed (Treagust, 1988), and the suitability between the student's answer choices and the reason is easy to know (Diani et al., 2019).

To reduce the weakness of the closed polytomous response test, the experts modified the test into an open polytomous response test. The open polytomous response test is a form of multiple-choice test that provides a place to write arguments for the answer choices (Retnawati, 2014). The studies on the open polytomous response test that have been carried out are test on calculus material in universities (Yang et al., 2017) and test on mathematics material in senior schools (Ayanwale, 2021) and junior schools (Falani et al., 2020). These studies developed open polytomous response tests for students in college and senior or junior school. Students in college and senior or junior school learn mathematics as a primary subject, while students in vocational schools learn mathematics as a secondary subject (Oktaria, 2016). In addition, students in vocational schools are more oriented towards practical abilities and skills, in contrast to students in college or high school who are more academically oriented, including in Indonesia (Permendikbud, 2016).

Currently, the Indonesian government expects that vocational schools are not academically left behind especially in mathematics. The government's commitment is to improve the method of assessing student learning in vocational school, and the current assessment method is using polytomous test. Often, students pay less attention during math exams for several reasons, such as considering mathematics as an unimportant subject (Putri et al., 2017), mathematics as a complicated subject (Vani et al., 2019), and mathematics as a boring subject (Ikmalwati, 2020). Therefore, the students tend to answer the test by guessing. To avoid the tendencies, it is necessary to develop a polytomous response test (closed or open). By considering the disadvantages of the closed polytomous response test, it is reasonable to conduct research by developing open polytomous response test for students in vocational schools.

The test instrument developed must be reliable as a good test, and therefore it is necessary to analyze the quality of test items (Rosidin, 2017). There are two theories for analyzing the item quality, namely classical and modern. Classical theory is a measurement theory for assessing test based on the assumption of measurement errors between actual results and observations, and from the assumption, a formula for calculating the level of difficulty and item discrimination was developed (Hambleton & Jones, 1993). The modern theory is a measurement theory to assess students' abilities by comparing students' abilities with their group abilities, and it is known as Item Response Theory/IRT (Hambleton & Linden, 1982). Classical theory is widely used by teachers because it is easy to apply. However, this classical theory has a weakness, namely, it cannot separate the characteristics of students and items. The modern theory is a solution to overcome the weaknesses of the classical theory because, in the modern theory, an item does not affect other items (local independence), items only measure one dimension (unidimensional) (Anisa, 2013), and an item eliminates the relationship between respondents and items (parameter invariance) (Saepuzaman et al., 2021). Therefore, experts suggest that the test instrument is accountable; the quality of the items must be good according to the analysis of classical and modern theory (Retnawati, 2014).

Therefore, the aim of the research is to develop a good mathematical test instrument using polytomous responses according to classical and modern theories in vocational schools. The research problems are stated as follows: (1) Does

the open polytomous response test developed have a good category so that it can be used as an assessment instrument in vocational schools based on classical and modern theory? and (2) Does the open polytomous response test instrument developed provide information on students' actual competence in vocational schools?

Methodology

Research Design

This research is a research and development model that refers to Plomp's (2013) model, with the research procedure consisting of five stages: preliminary investigation, design, realization or construction, test phase, revision, and implementation (test).

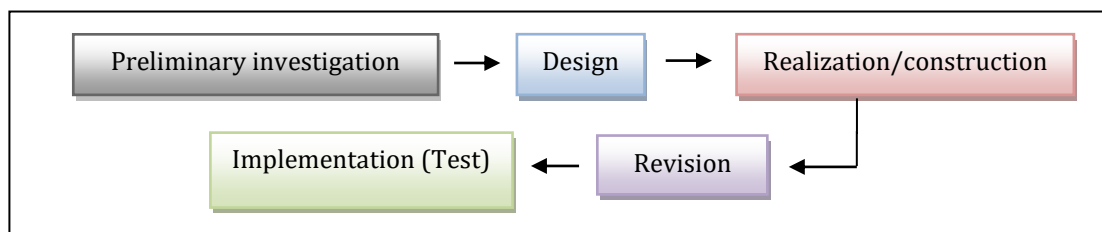


Figure 2. Stages of Research Design

The preliminary investigation stage is to identify the current assessment instruments used by teachers. The design stage is to make the open polytomous response test grid according to the basic competencies of mathematical concepts and to make a quality assessment questionnaire sheet. The realization/construction stage is developing the items test validated by an expert process for the items test. The revision stage is the improvement of items on the test based on expert advice. The implementation (testing) stage is to try out the test on students and to analyze the results of the test.

Research Subject

The subjects of the study were students of vocational schools situated in the province of Lampung, Indonesia. The research sample was determined using a non-probability sampling technique in the form of accidental sampling, which means taking a subject based on a subject that is easy to find and ready to be a respondent (Malhotra, 2006). The selected schools were three schools as representatives, namely the Mitra Bhakti vocational school, the Praja Utama vocational school, and the Ma'arif NU vocational school. The research subjects were 413 students in grade I (male students = 191, and female students = 222), whose mathematical abilities on the National Exam (NE) were categorized as moderate (average 64.67 out of the ideal score of 100). The following are the characteristics of the research subjects in detail shown in Table 1.

Table 1. Research Subjects

Vocational Schools	Grade	The Number of Students			Average of NE
		Total	Male	Female	
Mitra Bhakti	I-A	13	5	8	64.24
	I-B	23	10	13	62.4
	I-C	29	9	20	65
Praja Utama	I-A	40	40	0	64
	I-B	39	39	0	64.62
	I-C	43	15	28	68.72
	I-D	38	15	23	67.9
	I-E	41	14	27	67.56
	I-F	40	17	23	63.38
	I-G	40	8	23	67.74
Ma'arif NU	I-A	31	14	17	62.1
	I-B	36	5	51	58.34
Total		413	191	222	64.67

Data Collection Techniques

Data were collected using a questionnaire and a test. The questionnaire contains several questions to the teacher about the instrument for assessment used by the teacher and also expert validity of the instrument developed to determine content validity (Suhaini et al., 2021). The instrument was validated by two raters who have expertise in mathematics and educational evaluation. The three aspects assessed by the expert were the suitability of the items with the

indicators, language, and alternative answers to the questions. The score validating the instrument follows the criteria as in Table 2 below.

Table 2. Content Validity Score

Description	Score
Irrelevant	1
Quite relevant	2
Relevant	3
Very relevant	4

After determining the content validity, the instrument was tested to students. Then, it was continued by determining the validity of the construct and its reliability to ensure that the instrument could be further analyzed.

The instrument used was the open polytomous response test, which consisted of 40 questions on the concepts of sequences and series (arithmetic and geometry), quadratic equations and functions, and matrices. Each item contained five answer choices along with the reasons. Student scores are referred to the polytomous score in the Partial Credit Model, where answer choices and reasons were related (Retnawati, 2014), as shown in Table 3 below.

Table 3. Scores for Student Answers

Student Answers		Score
Answer Options	Reason	
False	False	1
False	Right	2
Right	False	3
Right	Right	4

Data Analysis

The collected data were analyzed in two stages: (1) questionnaire data analysis (qualitative analysis) and (2) test data analysis (empirical analysis). The following is an explanation of each data analysis:

1. Questionnaire data analysis (qualitative analysis)

There are two sets of questionnaire data namely, identification of assessment instruments in schools and expert assessment of the assessment instruments developed. The results of the two questionnaires were analyzed descriptively. Specifically, for expert judgment, it was continued with an analysis of expert agreement that used the Gregory index formula (Gregory, 2015), namely:

$$V = \frac{D}{A+B+C+D}$$

Description:

V = Content Validity

A = Rater 1 and 2 are weak

B = Rater 1 is strong, and rater 2 is weak

C = Rater 1 is weak, and rater 2 is strong

D = Rater 1 and 2 are strong

The interpretation of Gregory's formula is that the number V is in the range of 0 to 1. The higher the number V (close to 1 or equal to 1), the higher the value of the validity of an item. Conversely, the lower the number V (close to 0 or equal to 0), the lower the validity of an item.

2. Test data analysis (empirical analysis)

After conducting the content validity test, the researcher conducted the construct validity and reliability test. The construct validation test used exploratory factor analysis. The instrument is considered to have good construct if the explained Kaiser-Meyer-Olkin (KMO) value is greater than 0.5 (Retnawati, 2014). Reliability test using Cronbach's alpha formula. The instrument is said to have good reliability if the coefficient value of Cronbach's alpha is 0.60 (Arikunto, 2012). If the instrument has good construct validation, further test can be analyzed, namely the level of difficulty and item discrimination. The reason for analyzing the level of difficulty and item discrimination is that they are both preliminary analyses of the assumptions of measurement theory (Hambleton & Jones, 1993). To simplify the process of analyzing the level of difficulty and item discrimination, the Iteman program was used for classical theory

and the Winsteps program for modern theory (Sarea & Ruslan, 2019). The Winsteps program was used because it had several advantages; namely, it can analyze polytomous data and calculate the maximum likelihood model using a 1-parameter logistic model (Untary et al., 2020).

2.1. Analysis of test data with classical theory

- a. The item difficulty level is the percentage of the number of students who answered correctly or incorrectly. If the item has an index of 0.3-0.7, then the item is good; if the item has an index below 0.3, then the item is difficult; and if the item has an index above 0.7, then the item is easy.
- b. Discrimination is the ability of a test to distinguish between high-ability students and low-ability students. Discrimination is said to be good if it has an index above 0.3, and if the discrimination index is below 0.3, then the question needs to be revised (Arikunto, 2012).

2.2. Analysis of test data with modern theory

- a. The item difficulty level is the level of the student's latent trait towards the item. The difficulty of the items determines the ability of about 50% of the respondents who are expected to answer items correctly (DeMars, 2010). An item is said to be good if it has an index of between -2 and +2 (Hambleton & Swaminathan, 1985). If the index is close to -2, then the item is classified as very easy, and if the index is close to +2, the item is classified as very difficult (Retnawati, 2014). In the Winsteps program, the item difficulty level is in the Measure column.
- b. Item discrimination is indicated by the slope of the curve on the item characteristics. The item is said to have good discrimination if the slope of the curve is moderate (not too gentle or steep) because if the slope of the curve is too gentle or steep, then the item is not good. Another opinion states that a good index is above 0.4 (Crocker & Algina, 1986). In the Winsteps program, the item discrimination is in the Pt-Measure Correlation column.

According to modern theory, before analyzing the item difficulty level and discrimination, three assumptions must be tested, namely unidimensionality, local independence, and model fit (Hambleton et al., 1991). Unidimensional means that each test item only measures one ability. There are three ways that are often used to test unidimensionality, namely the analysis of the Eigenvalue of the correlation matrix between items, the Stout-test on the unidimensional assumption test, and the index based on the residuals of the unidimensional solution (DeMars, 2010). In this study, the dimensional test used the Eigenvalue analysis of the correlation matrix between items.

Local independence is the state of the respondent's answer to an item that is not influenced by other items. Local independence test by proving that the probability of the respondent's answer pattern is the same as the probability of the respondent's answer to each item. If the unidimensional assumption is accepted, the local independence assumption will also be accepted (DeMars, 2010). Use the Model Fit Test to find out whether the model used is in accordance with the items. Test the fit of the model by measuring the outfit mean square (MNSQ) and PT-Measure. If the outfit's MNSQ value is 0.5 to 1.5 and the Pt-Measure Correlation is positive, it is said that the item fits the model (Linacre, 2012). In addition, the information function and standard error measurement (SEM) are analyzed, which aims to further explain the latent ability as measured by using a test that is expressed through item donations.

Results

Analysis of Questionnaire Data

Based on the results of the questionnaire, it was found that the teacher had never used the polytomous response. As many as 80% of teachers used essay tests and 20% of teachers used multiple-choice tests, with each instrument consisting of 2-5 items. In addition, about 10% of teachers used this assessment as a means for learning improvement, such as improving lesson plans and teaching methods. The results of the questionnaire stated that 90% of teachers who did not use assessment as an improvement in learning were caused by several aspects, such as teachers did not understand assessment (20%), teachers did not know how to analyze assessments (50%), and teachers did not know how to develop good assessment questions (30%). The following is the summary of the questionnaire from the identification data.

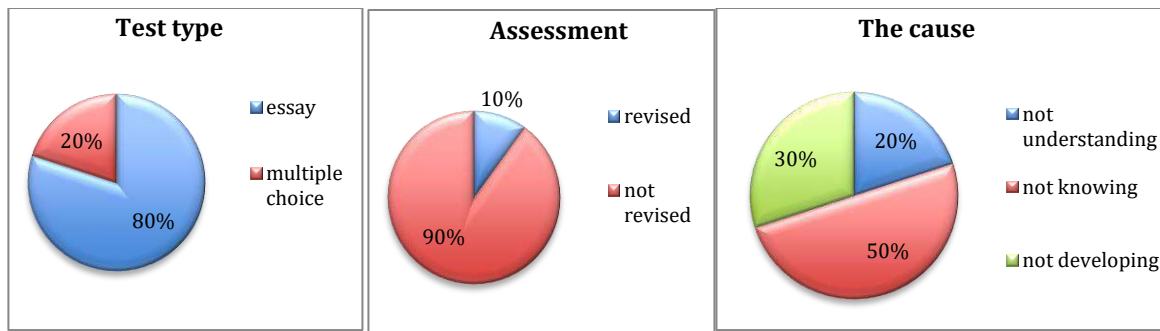


Figure 3. Description of Teacher Conditions in Assessment

Content Validity

The results of the two expert assessments showed that the content validation instrument is good. Furthermore, the analysis of expert judgment agreement was obtained as shown in Table 4 below.

Table 4. The Gregory Item Index

		Rater 1	
		Weak	strong
Rater 2	weak	0	0
	strong	0	40
Index Gregory		1	

Based on the results of the assessment in Table 4, it can be concluded that the instrument is valid because the value of V reaches a value of 1. Therefore, the instrument test can be continued. In addition to providing assessments, the experts also provided some suggestions for improvements to the instrument, namely the preparation of questions using the ABCD format (Audience, Behavior, Competence, and Degree), avoiding the use of ambiguous language or statements, improving mathematical concepts, making alternative answer choices that are misleading, and arranging them in order.

Analysis of Test Data

Construct Validity

After testing the instrument, it was followed by a construct validity test. The results of the test with exploratory factor analysis are shown in Table 5.

Table 5. Exploratory Factor Analysis

KMO and Bartlett's Test		Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	1936.378
	Df	780
	Sig.	.000

Based on Table 5, the explained KMO value is 0.76 (more than 0.5), and it can be concluded that the variables and samples used to allow for further analysis.

Reliability

The results of the estimation of the reliability of the instrument obtained a Cronbach's alpha coefficient value is 0.89 (more than 0.6). It means that the instrument has good reliability so that the analysis of the level of difficulty and item discrimination can be continued according to classical and modern methods.

Table 6. Reliability of Items

Cronbach's Alpha	N of Items
0.892	40

Analysis of Test Data with Classical Theory

Analysis of test data in the classical way did not require testing assumptions, but the analysis of the difficulty level of the items and the distinguishing power of items could be directly calculated if the validity and reliability have been met. The results of the two analyses are presented in Table 7 below.

Table 7. The Item Difficulty Level and Discrimination

Item	Difficulty Level	Category	Discrimination	Category	Item	Difficulty Level	Category	Discrimination	Category
1	0.538	Good	0.196	Revised	21	0.482	Good	0.143	Revised
2	0.487	Good	0.179	Revised	22	0.535	Good	0.429	Good
3	0.528	Good	0.214	Revised	23	0.492	Good	0.250	Revised
4	0.540	Good	0.304	Good	24	0.438	Good	-0.071	Revised
5	0.489	Good	0.089	Revised	25	0.436	Good	-0.107	Revised
6	0.446	Good	-0.161	Revised	26	0.385	Good	-0.286	Revised
7	0.438	Good	-0.232	Revised	27	0.383	Good	-0.321	Revised
8	0.453	Good	-0.143	Revised	28	0.416	Good	-0.143	Revised
9	0.414	Good	-0.143	Revised	29	0.458	Good	-0.125	Revised
10	0.409	Good	-0.339	Revised	30	0.385	Good	-0.375	Revised
11	0.438	Good	-0.143	Revised	31	0.404	Good	-0.321	Revised
12	0.436	Good	-0.036	Revised	32	0.433	Good	-0.250	Revised
13	0.400	Good	-0.321	Revised	33	0.441	Good	0,036	Revised
14	0.450	Good	-0.036	Revised	34	0.424	Good	-0.268	Revised
15	0.462	Good	0.250	Revised	35	0.412	Good	-0.321	Revised
16	0.453	Good	-0.089	Revised	36	0.431	Good	-0.304	Revised
17	0.416	Good	-0.143	Revised	37	0.404	Good	-0.232	Revised
18	0.419	Good	-0.196	Revised	38	0.363	Good	-0.482	Revised
19	0.431	Good	-0.232	Revised	39	0.230	Good	-0.929	Revised
20	0.441	Good	-0.089	Revised	40	0.211	Good	-1.071	Revised

Based on Table 7, it was found that all items have an item difficulty level in the index range of 0.3 to 0.7, so they are categorized in the good category. Meanwhile, only two items on discrimination had good categories, and the remaining items needed to be revised. The results indicated that all items were good based on the level of difficulty, but almost all items needed to be revised for item discrimination.

*Analysis of Test Data with Modern Theory**The Unidimensional Assumption Test*

The unidimensional assumption test is the first assumption test with factor analysis. Factor analysis begins by testing the adequacy of the sample to be used in the analysis, constructing a variance-covariance matrix, and then calculating the Eigenvalue. The Eigenvalue was then used to calculate the percentage of explained variance, as well as to describe the scree plot (Retnawati, 2014). The output of factor analysis was the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) statistic and scree plot.

Table 8. The KMO Test

KMO Test	Score
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.760
Sig.	.000

The unidimensional test is seen based on the cumulative percentage of Eigenvalue and scree plot analysis results. If the cumulative percentage of the first-factor Eigenvalue is greater than 20%, then the unidimensional assumption is fulfilled (Retnawati, 2014). In Table 9, it can be seen that the cumulative percentage of the first-factor Eigenvalue is 20.220%. Because the Eigenvalue is more than 20%, this instrument is proven to only measure one factor or dimension.

Table 9. Total Variance Explained

Component	Initial Eigenvalue		
	Total	Variance (%)	Cumulative (%)
1	8.088	20.220	20.220
2	1.458	3.646	23.865

In addition, the unidimensional test can also be seen on the scree plot, which is based on the number of factors marked by the steepness of the graph with the acquisition of Eigenvalue.

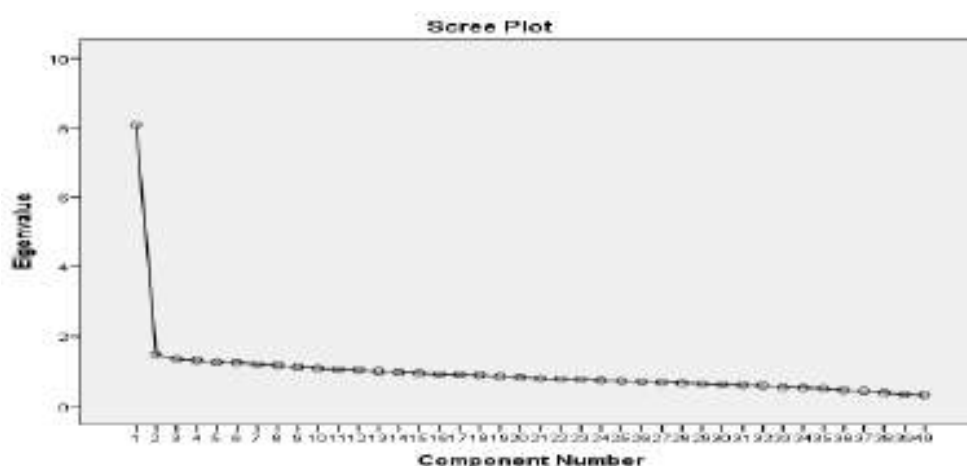


Figure 4. Scree Plot Unidimensional

Based on the scree plot, it is known that the Eigenvalue immediately slope on the second factor. It demonstrates that the developed instrument has only one dominant factor. The results prove that the test kit meets the unidimensional assumption, or in other words, only measures one dominant factor.

Local Independence Assumption Test

The local independence assumption test will be fulfilled if the student's answer to one item does not affect the student's answer to another item. Thus, the score of one item should not be determined or dependent on the scores of other items. This confirms that this assumption automatically proves that students' answers do not affect answers to other items (Retnawati, 2014).

Table 10. The Covariance Matrix.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.0486									
K2	0.012	0.0042								
K3	0.0105	0.0034	0.0032							
K4	0.0317	0.0106	0.0093	0.0293						
K5	0.01	0.0032	0.0029	0.0085	0.0028					
K6	0.0069	0.0023	0.0021	0.0062	0.002	0.0016				
K7	0.0036	0.0011	0.0009	0.0028	0.0009	0.0007	0.0004			
K8	0.0041	0.0015	0.0012	0.0037	0.0011	0.0008	0.0004	0.0006		
K9	0.0074	0.0024	0.0022	0.0065	0.002	0.0015	0.0007	0.0009	0.0017	
K10	0.021	0.0061	0.0054	0.0148	0.0051	0.0039	0.0017	0.0025	0.0043	0.0196

Table 10 shows the results of the variance-covariance values between groups of students' abilities. In the table, it can be seen that the covariance value between the ability interval groups located on the diagonal is small and close to zero. This result shows that there is no correlation, so it can be said that the local independence assumption test is accepted.

Model Fit

The model fit test was analyzed using the Winsteps program. The item requirements are called "fit to the model." If the Outfit MNSQ value is 0.5 to 1.5 and the Outfit ZSTD value is -2 to 2, and the Pt-Measure Correlation is positive, then it can be said that the item fits the model (Sumintono & Widhiarso, 2015). An item is considered fit if one of the conditions is accepted. In addition, it can also be seen from the MNSQ infit of 0.77 to 1.3, but at this stage, the fit of the model is only taken on the MNSQ and Pt-Measure outfit values. Based on the results of the analysis, all items matched the model or fit (Figure 5).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	-.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	-.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 5. Item Fit on Model

The Item Difficulty Level

The item difficulty level was analyzed using the Winsteps program, and the results obtained were presented in Table 12 (Measure column). The item difficulty level is in the range of -0.70 to 0.84. From Table 12, the highest difficulty level is item 40 (difficulty level 0.84), and the lowest difficulty level is item 1 (difficulty level -0.70). Since the difficulty level is in the range of 2 and 2, it can be concluded that all items are in the good category. If further divided into three categories, the difficulty level of items in the range of -0.7 to 0.84 is moderate (Sumintono & Widhiarso, 2015). It can also be seen on the difficult map items, namely that the difficulty level is in the range of -2 and 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
1	1134	413	-0.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-0.45	.07	.97	-0.4	.98	-0.3	.75	.43	49.2	49.1	Q2
3	1066	413	-0.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-0.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-0.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-0.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.05	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	-.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	-.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 6. Item Difficulty Level

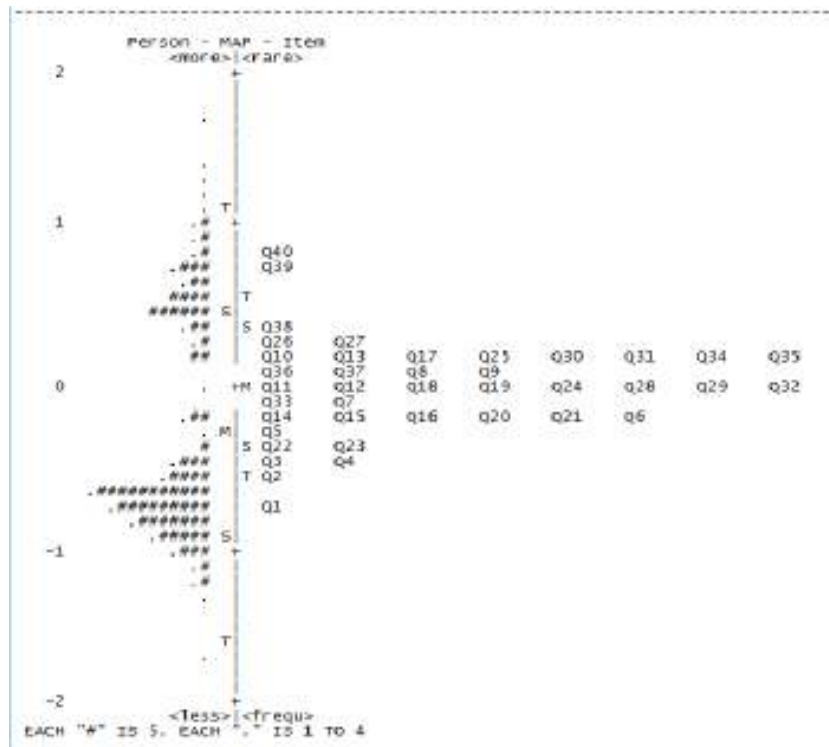


Figure 7. Item Difficulty Map

The Item Discrimination

The item discrimination analysis used the Winsteps program, and the results are presented in Figure 8 (PT-Measure Correlation column). Item discrimination is in the range of 0.23 to 0.74, with details of 27 items having an index above 0.4 (good category), and 13 items having an index below 0.4 (bad category). However, the 13 items can be used as an instrument as long as the index is above 0 (Alagumalai et al., 2005).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	Item
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.			
1	1134	413	-.70	.07	1.16	2.5	1.16	2.5	.74	.43	35.1	47.9	Q1
2	1079	413	-.45	.07	.97	-.4	.98	-.3	.75	.43	49.2	49.1	Q2
3	1066	413	-.39	.07	.93	-1.0	.94	-1.0	.67	.43	49.4	49.5	Q3
4	1064	413	-.38	.07	.82	-3.0	.82	-3.0	.46	.43	50.1	49.6	Q4
5	1021	413	-.19	.07	.84	-2.6	.84	-2.6	.41	.44	53.5	50.6	Q5
6	998	413	-.09	.07	1.16	2.4	1.16	2.4	.44	.44	42.4	50.9	Q6
7	977	413	.01	.07	1.03	.5	1.03	.5	.45	.44	40.7	51.1	Q7
8	966	413	.06	.07	1.16	2.4	1.16	2.5	.51	.44	42.1	51.1	Q8
9	965	413	.06	.07	.97	-.5	.97	-.5	.43	.44	49.9	51.0	Q9
10	964	413	.07	.07	.93	-1.0	.93	-1.0	.40	.44	49.2	51.0	Q10
11	986	413	-.03	.07	.97	-.4	.98	-.4	.38	.44	44.1	51.0	Q11
12	988	413	-.04	.07	.90	-1.7	.90	-1.7	.39	.44	52.3	51.0	Q12
13	957	413	-.10	.07	1.07	1.1	1.07	1.1	.44	.44	47.5	51.0	Q13
14	1003	413	-.11	.07	1.01	.2	1.01	.3	.42	.44	48.7	50.9	Q14
15	1000	413	-.09	.07	1.10	1.5	1.10	1.6	.49	.44	42.9	50.9	Q15
16	994	413	-.07	.07	1.08	1.2	1.08	1.2	.44	.44	46.5	51.0	Q16
17	964	413	.07	.07	.96	-.7	.96	-.6	.48	.44	52.8	51.0	Q17
18	972	413	.03	.07	1.12	1.8	1.12	1.8	.35	.44	43.1	51.1	Q18
19	974	413	.02	.07	1.04	.7	1.05	.8	.26	.44	47.7	51.1	Q19
20	997	413	-.08	.07	.97	-.4	.97	-.4	.47	.44	50.8	50.9	Q20
21	1010	413	-.14	.07	1.01	.3	1.02	.3	.44	.44	47.5	50.8	Q21
22	1050	413	-.32	.07	1.08	1.2	1.08	1.2	.56	.44	43.6	50.0	Q22
23	1037	413	-.26	.07	1.05	.8	1.05	.8	.53	.44	47.2	50.3	Q23
24	974	413	.02	.07	1.06	1.0	1.07	1.1	.37	.44	46.7	51.1	Q24
25	968	413	.05	.07	1.00	.0	1.00	.0	.41	.44	47.9	51.1	Q25
26	943	413	.16	.07	.97	-.4	.97	-.4	.31	.44	50.8	50.9	Q26
27	936	413	.20	.07	1.02	.3	1.02	.4	.32	.44	46.0	50.8	Q27
28	975	413	.02	.07	.89	-1.8	.88	-1.9	.40	.44	54.5	51.1	Q28
29	981	413	-.01	.07	.97	-.5	.97	-.5	.33	.44	49.2	51.0	Q29
30	949	413	.14	.07	1.06	.9	1.05	.9	.30	.44	45.5	50.9	Q30
31	955	413	.11	.07	1.07	1.1	1.07	1.1	.23	.44	44.8	51.0	Q31
32	974	413	.02	.07	1.12	1.8	1.12	1.8	.35	.44	43.3	51.1	Q32
33	981	413	-.01	.07	1.04	.6	1.04	.6	.41	.44	48.2	51.0	Q33
34	959	413	.09	.07	1.06	1.0	1.07	1.1	.33	.44	46.5	51.0	Q34
35	955	413	.11	.07	1.04	.7	1.04	.7	.40	.44	47.2	51.0	Q35
36	961	413	.08	.07	1.04	.6	1.05	.8	.34	.44	47.7	51.0	Q36
37	955	413	.11	.07	.96	-.6	.96	-.6	.40	.44	51.1	51.0	Q37
38	921	413	.27	.07	.89	-1.7	.89	-1.8	.43	.44	52.8	50.6	Q38
39	825	413	.72	.07	.65	-6.3	.66	-6.2	.59	.43	58.1	48.6	Q39
40	801	413	.84	.07	.74	-4.5	.74	-4.6	.54	.43	55.4	48.2	Q40
MEAN	979.5	413.0	.00	.07	1.00	-.1	1.00	-.1			47.8	50.6	
S. D.	56.8	.0	.26	.00	.11	1.8	.11	1.8			4.3	.8	

Figure 8. Item Discrimination

Comparative Analysis Between Classical and Modern Theory

The results of the analysis of classical and modern theories obtained the index of difficulty level and item discrimination as follows.

Table 11. Comparison of Classical and Modern Theories

Parameter	Classical Theory		Modern Theory	
	Many Items with Good Categories	Percentage	Many Items with Good Categories	Percentage
Difficulty level	40	100	40	100
Item discrimination	2	0.05	27	67.5

Based on Table 11, the level of item difficulty analyzed by classical and modern theory has the same results (good category). However, item discrimination using modern theory has more items in the good category than classical theory. If we compare the index of discriminatory items between classical (Table 7) and modern (Figures 6 and 8), it can be seen that there is a match between the categories of item discrimination. It means that if item discrimination is not good with the classical theory, then item discrimination is also not good with the modern theory (13 items correspond to the item discrimination index, and 27 do not match).

Information and Measurement Error (SEM) Function

The function of information to reveal latent ability was measured by using a test that was expressed through item donation. The test information function is also the sum of the functions of each item. The information function is inversely proportional to measurement error, or standard error measurement (SEM). The value of the information function of the test device will be high if the items that make up the test have a high information function. The following is a picture of the curve of the relationship between the information function and SEM.

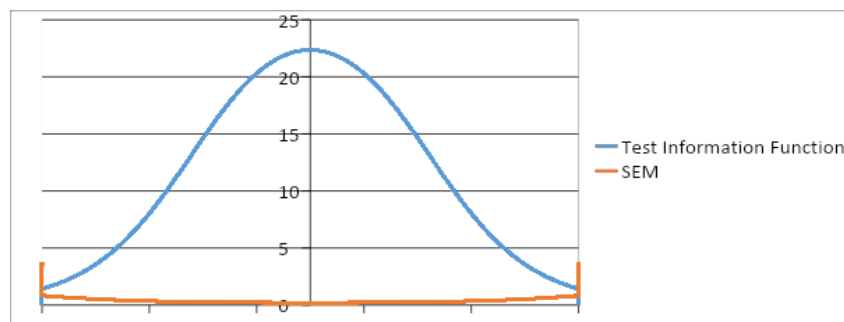


Figure 9. Graph of Information Function and Measurement Error

Figure 9 shows that this instrument provides information of 22.36 (maximum) and has a measurement error of 0.21 (smallest) for medium-ability students. The lower and upper limits of the interval are the ability scores where the graph of the information function and the SEM graph intersect at that interval. This graph states that the greater the value of the information function, the smaller the measurement error (SEM), and the item information function expresses the strength or contribution of the test items in revealing the latent trait as measured by the test. This information function provides a description of the item according to the model (which helps with item selection) (Retnawati, 2014). These results conclude that this test instrument is suitable for students with medium abilities.

Based on the test results, it can also be seen that the actual competency information of students in the vocational school was based on the test answers. Instrument analysis is based on two patterns of student answers that have the same tendency based on Bloom's Taxonomy (Bloom, 1956). A total of six student answers were selected as samples with different abilities (high, medium, and low).

Item 1: The cognitive domain to be achieved is C2 (Understanding). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood the general form of arithmetic sequences and know the first term and other terms, and (2) students have not been able to formulate general forms in arithmetic sequences and perform algebraic operations on general forms arithmetic sequence.

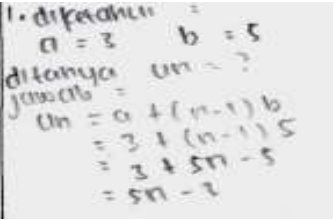
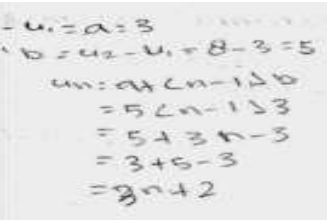
<p>Question 1:</p> <p>Given an arithmetic sequence: 3, 8, 13, 18, The formula for the nth term of the sequence is ...</p> <p>A. $U_n = 5n - 3$ B. $U_n = 5n - 2$ C. $U_n = 2n + 1$ D. $U_n = 4n - 1$ E. $U_n = 3n + 2$</p> <p>Reason:</p>	<p>Pattern 1:</p> 	<p>Pattern 2:</p> 
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------

Figure 10. An Example of Student Answers in Item 1

Item 2: The cognitive domain to be achieved is C2 (understanding). Based on students' answers, the two dominant patterns of student answers are (1) being able to understand the number of terms in a sequence by using the general formula for an arithmetic sequence, or determining the number of arithmetic sequences without using a general formula (only writing down all the terms from the first term until the last term) and (2) students who can already use the general formula for arithmetic sequences but have not been able to determine the number of arithmetic sequences because of errors in performing algebraic operations on general arithmetic sequences.

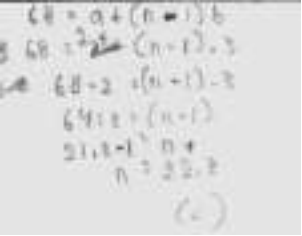
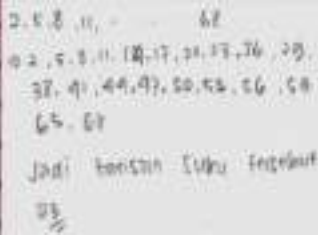
<p>Question 2:</p> <p>Given an arithmetic sequence: 2, 5, 8, 11, ..., 68. The number of terms in the sequence is...</p> <p>A. 12 B. 13 C. 22 D. 23 E. 24</p> <p>Reason:</p>	<p>Pattern 1</p> 	<p>Pattern 2</p> 
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------

Figure 11. An Example of Student Answers in Item 2

Item 3: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the difference between two non-adjacent arithmetic sequences using the general formula for arithmetic sequences, and (2) other students can determine the number of arithmetic sequences even though they are not using a general formula or by writing the terms of the known terms and inserting several terms.

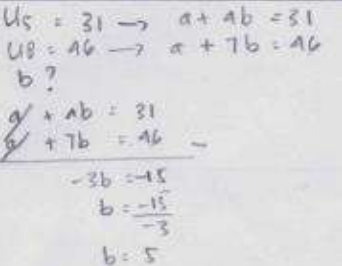
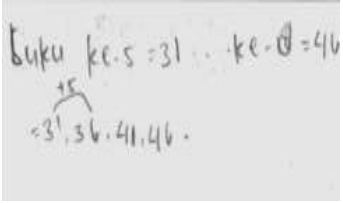
<p>Question 3:</p> <p>An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 B. 6 C. 7 D. 8 E. 11</p> <p>Reason:</p>	<p>Pattern 1</p> 	<p>Pattern 2</p> 
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------

Figure 12. An Example of Student Answers in Item 3

Item 4: The cognitive domain to be achieved is C3 (applying). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and can determine the nth term in an arithmetic sequence that is known to be two non-adjacent terms using the general formula for arithmetic sequences, and (2) students cannot determine the number of arithmetic sequences because it does not use a general formula but by writing the terms of the known terms and inserting several terms and continuing until the nth term.

Question 4:
 The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...
 A. 308
 B. 318
 C. 326
 D. 344
 E. 354

Reason:

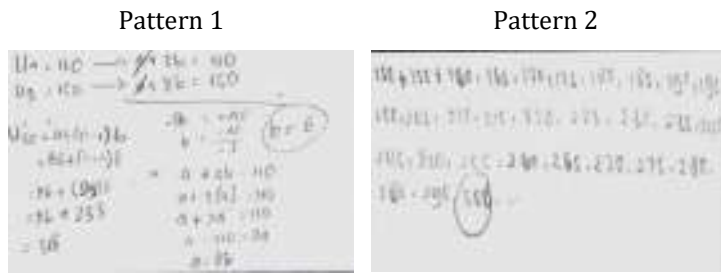


Figure 13. An Example of Student Answers in Item 4

Item 5: The cognitive domain to be achieved is C1 (remembering). Based on students' answers, the two dominant patterns of student answers are: (1) students have understood and determined the middle term of an arithmetic sequence, and (2) students can determine the middle term of an arithmetic sequence but do not use general formulas, but rather by writing the terms from known terms and inserting several terms and then defining them.

Question 5:
 An arithmetic sequence is: 3, 8, 13, 18, ..., 103.
 Then the middle term of the sequence is ...308
 A. 53
 B. 52
 C. 20
 D. 11
 E. 10

Reason:

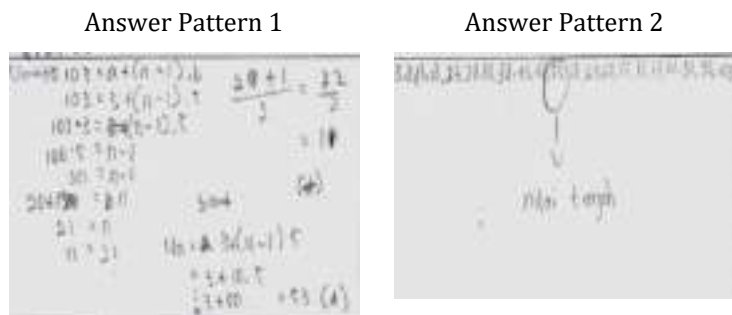


Figure 14. An Example of Student Answers in Item 5

Description of Instrument Development and Student Ability

The instrument developed from this research is the open polytomous response test, and all parameters have been accepted. This instrument is a combination of a multiple-choice test and an essay. Multiple choice test is easier to check students' answers, but their mathematical thinking processes cannot be known in depth. While the essay test has the advantage of being able to find deeper mathematical thinking processes, it takes a long time to check the answers.

Analysis of learning instruments is an important source of composite scores in the final report. In the final report, the student's ability score was first changed to a score of 0–10 (previously 0–100). The conversion is done through a linear transformation by dividing the student's score by the ideal score. Then, the result is multiplied by 10 to get a value range of 0–10 or multiplied by 100 to get a value range of 0–100. In the range of 0–10, the scores of students' mathematical ability were 8.56 (highest) and 4.31 (lowest), or in the range of 0–100, students' math ability scores were 85.6 (highest) and 43.1 (lowest).

The results of the analysis of student abilities are presented in the form of predicates ranging from very low to very high according to the specified category. The results of this analysis show that most students have very low to medium abilities, as much as 62% (253 students), and the remaining 38% (160 students) have high and very high abilities. Other analysis results found that high and very high-ability students tend to work according to concepts with more creative completion steps (different from the teacher's example), but students who have very low to medium abilities can solve problems according to concepts with less creative completion steps (e.g., routine or according to the teacher's example).

Other results show that the assessment with an open response polytomous makes it easier for teachers to explore students' difficulties with a material. Then, from this exploration, the teacher can continue with improvements for students who have learning difficulties. An important finding of this study is that information about students' learning difficulties through the open polytomous response test is more secondhand and complete than other assessment instruments, such as multiple-choice tests(Gierl et al., 2017) or essay tests(Putri et al., 2020).

Discussion

This research is development research aimed at developing a good mathematical assessment instrument using polytomous responses according to classical and modern theories. The results of the data analysis found that there were differences in the results of the analysis between classical theory and modern theory, namely on item

discrimination. Classical theory analysis obtained 38 items with bad criteria and only 2 items with good criteria. In contrast, modern theory analysis obtained 27 items with good criteria and 13 items with bad criteria. According to evaluation theory, modern theory aims to cover the weaknesses of classical theory. The results of the classical theory analysis are often categorized as poor, but modern analysis results are categorized as good, and vice versa (Retnawati, 2014). That is, an item that is not in a good category with classical theory should be analyzed according to the modern theory before revising or replacing the item.

Research on learning assessment with open response polytomous was carried out by Yang et al. (2017). This research aims to diagnose student errors in completing calculus material at university. The instrument compares two types of test namely the two-tier test and the open polytomous response test. The research findings suggest that the open polytomous response test provides more detailed information on student error than the two-level test (see Figure 15). The research findings are in line with the results of this study, which states that the open polytomous response test provides more detailed information about students' abilities.

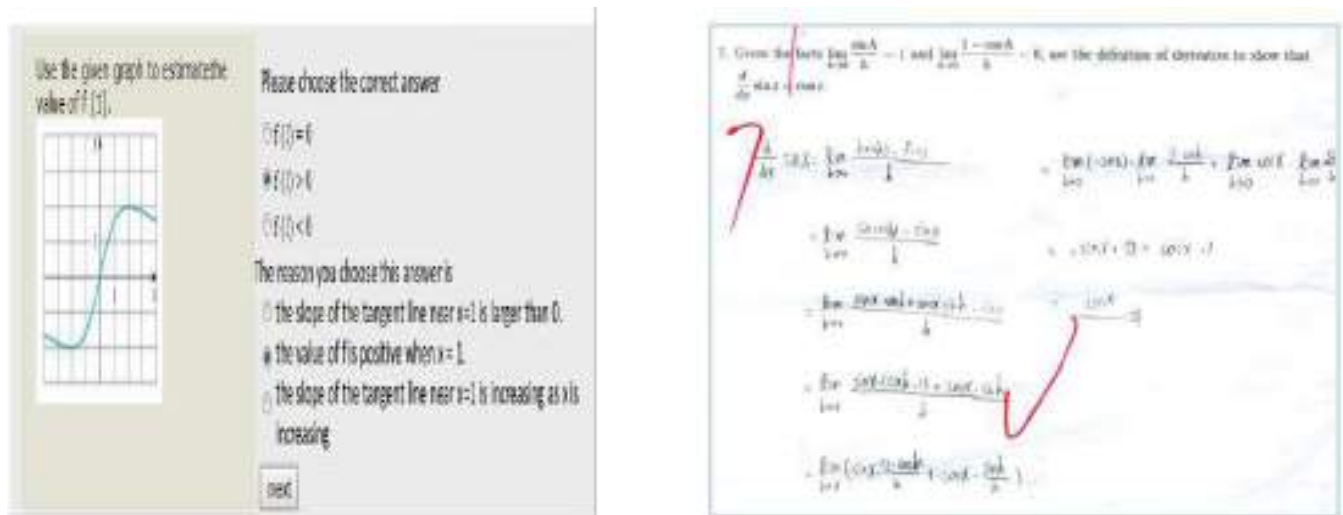


Figure 15. An Example of Student Answers on The Two-Tier Test and The Open Polytomous

Response Test

Another study on learning assessment with open response polychromous was conducted by Ayanwale(2021).Ayanwale's research compares two polytomous response test analysis methods namely the Parallel and Partial Credit Model. The results of his research stated that the Partial Credit Model analysis had a first factor Eigenvalue of 20.5% (ideal Eigenvalue of more than 20%) compared to the Parallel analysis, which had the first factor Eigenvalue of 11.7%. The results of Ayanwale's research are in line with this study, which obtained the first factor Eigenvalue of 20.220%, and the author can state that the instrument developed is suitable to be used to assess vocational students in Indonesia, maybe even outside Indonesia.

Other studies related to classical and modern theory were conducted bySarea (2018)and Saepuzaman et al.(2021). Sarea's research states that the response polytomous test has good criteria (classical and modern theory), and the classical theory analysis has more items than modern theory. Meanwhile, Saepuzaman's research found that the response polytomous test had good criteria (classical and modern theory), and the modern theory analysis had more items than classical theory.

The results of the analysis of student answers obtained information that there were two patterns of students' solving questions: (1) formulas and (2) trial and error. Students who use formula patterns in solving problems tend to be carried out by high-ability students, and students who use trial and error patterns tend to be carried out by students with medium and low abilities. Students who use both patterns can answer the questions correctly as shown in Figure 16 below.

The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...

- A. 308
- B. 318
- C. 326
- D. 344
- E. 354

Reason:

$$\begin{aligned}
 U_4 &= 110 \rightarrow a + 3b = 110 \\
 U_9 &= 150 \rightarrow a + 8b = 150 \\
 \hline
 & -5b = -40 \\
 & b = \frac{-40}{-5} \quad (b = 8) \\
 & a + 3(8) = 110 \\
 & a + 24 = 110 \\
 & a = 110 - 24 \\
 & a = 86 \\
 U_{30} &= a + (n-1)b \\
 &= 86 + (29)b \\
 &= 86 + 233 \\
 &= 318
 \end{aligned}$$

(i)

(ii)

Figure 16. An Example of Student Answer Patterns with (i) Formulas, (ii) Trial and Error

Both patterns of solving math problems are common ways. This is in line with the opinion of Mason et al. (2010) that there are several ways that can be used in solving math problems, namely trial and error, using a drawing or model, analogy, and formula. Syahlan(2017) states that two ways that students often use to solve math problems are (1) trial and error and (2) formulas. Usually, students use trial and error methods to solve easy problems and formulas methods to solve difficult questions. If combined with student answers and Syahlan's opinion, students with medium and low abilities in solving easy questions will use trial and error methods, and students with high abilities in solving difficult questions will use formulas.

The results of these student answers can be important information for teachers in teaching mathematics. This means that before teaching material, the teacher must know the students' initial abilities (high, medium, or low) and the level of difficulty of the material. Some of the benefits for teachers who know students' initial abilities are: teachers can develop professionally centered on students (Gonzalez, 2018), teachers can adjust the level of cognitive and increase student learning engagement (Dong et al., 2020), assist teachers in designing pedagogical practices and correcting students' misconceptions (Geofrey, 2021). In addition, the benefit for teachers who know the level of difficulty of a material is that they can design remedial learning plans (Muhson et al., 2017; Wulanningtyas et al., 2020). The description above shows that there are many benefits that the teacher will get if the teacher knows the students' actual abilities, and this can be known by the teacher if the teacher uses the open polytomous response test.

Conclusion

Based on the results of the research and discussion, it can be concluded that (1) the open polytomous response test has a good category according to classical and modern theory. Thus, the test instrument requirements are accountable (qualifies for a good test) is a good test instrument according to the analysis of the two theories (classical and modern theory), and (2) the open polytomous response test can provide information on the actual competence of students; this is observed in the students' arguments in giving reasons for their choices. This is observed in the students' arguments when giving reasons for their choices. Therefore, the open polytomous response test can be used as an alternative to learning assessment.

Recommendation

Based on the research results, there are several recommendations for teachers, schools, and other researchers. For teachers, they should familiarize students with giving a test in the form of a polytomous response before giving the test. In schools, principals or other leaders should encourage other teachers to take advantage of this test and develop other assessment instruments. For other researchers, they should conduct research by developing instruments with other polytomous responses (assessment of learning and assessment as learning) on other materials. In addition, for further research, it is suggested to conduct a study that develops an assessment instrument with a learning response polytomous (pretest). This is important so that students' prior knowledge can be known and learning can be effective.

Limitations

The research carried out has several limitations. Firstly, the selected schools have not met the researchers' expectations, for example, representing schools of high, medium, and low quality. In addition, during the collection of research data that begins with learning, it is not fully controlled by the researcher, so the students who are the research samples are less conditioned. The research is only limited to algebraic materials (sequences and series, matrices, equations, and quadratic functions), and does not represent other mathematical materials, such as geometry and statistics.

Authorship Contribution Statement

Sutiarso: Conceptualization, design, development of instruments, analysis, article writing, final approval. Rosidin: Design, development of instruments, analysis, editing. Sulistiawan: Development of instruments, collect data, analysis, editing.

References

- Alagumalai, S., Curtis, D., & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Springer-Kluwer Academic Publishers.
- Anisa. (2013). Perbandingan penskoran dikotomi dan politomi dalam teori respon butir untuk pengembangan bank soal mata kuliah matematika dasar [Comparison of scoring dichotomies and polytomies in item response theory for the development of a question bank for basic mathematics courses]. *Jurnal Matematika, Statistika, & Komputasi*, 9(2), 95-113. <https://bit.ly/39wv73P>
- Antara, A. A., Yasna, I. M., Dewi, N. W., & Maduriana, I. M. (2019). Karakteristik tes prestasi belajar model campuran dikotomis dan politomis generalized partial credit model [Characteristics of learning achievement test mixed dichotomous and polytomous generalized partial credit model]. *JurnalSuluh Pendidikan*, 17(1), 32-37. <https://bit.ly/3yFQ6eO>
- Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* [Educational evaluation basics]. Bumi Aksara.
- Ayanwale, M. A. (2021). Calibration of polytomous response mathematics achievement test using generalized partial credit model of item response theory. *Educatum: Journal of Science, Mathematics and Technology*, 8(1), 57-69. <https://bit.ly/3Nob4mN>
- Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.
- Chandrasegaran, A. L., Treagust, D., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307. <https://doi.org/10.1039/B7RP90006F>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.
- DeMars, C. (2010). *Item response theory*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>
- Diani, R., Alfin, J., Anggraeni, Y. M., Mustari, M., & Fujiani, D. (2019). Four-tier diagnostic test with certainty of response index on the concepts of fluid. In C. Anwar (Ed.), *Young scholar symposium on transdisciplinary in education and environment (YSSTEE)* (pp. 641-650). UIN Raden Intan. <https://bit.ly/3lfARKW>
- Dong, A., Jong, M. S., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontier Psychology Journal*, 11(1), 1-10. <https://doi.org/10.3389/fpsyg.2020.591203>
- Earl, L. M. (2013). *Assessment for learning; Assessment as learning: Changing practices means changing beliefs*. Assessment & Support Team. <https://bit.ly/3MnA9hw>
- Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of students' ability estimation on combinations of item response theory models. *International Journal of Instruction*, 13(4), 545-558. <https://doi.org/10.29333/iji.2020.13434a>
- Geoffrey, M. (2021). Children's prior knowledge is very important in teaching and learning in this era of constructivism. ResearchGate. <https://doi.org/10.13140/RG.2.2.28470.22083>
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice test in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. <https://doi.org/10.3102/0034654317726529>
- Gonzalez, G. (2018). Understanding teacher noticing of students' prior knowledge: Challenges and possibilities. *The Mathematics Enthusiast*, 15(3), 483-528. <https://doi.org/10.54870/1551-3440.1442>

- Gregory, R. J. (2015). *Psychological testing: History, principles, and applications*. Pearson.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378. <https://doi.org/10.1177/014662168200600401>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer. <https://doi.org/10.1007/978-94-017-1988-9>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal*, 25(1), 523–545. <https://doi.org/10.1007/s13394-013-0083-6>
- Ikmawati. (2020). Pengaruh disiplin dan kreativitas belajar terhadap hasil belajar matematika di SMK negeri dan swasta [The effect of discipline and creativity in learning on math learning outcomes in public and private vocational schools]. *Primatika: Jurnal Pendidikan Matematika*, 9(1), 35-42. <https://doi.org/10.30872/primatika.v9i1.250>
- Kartono. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar [Equalization of the dichotomous and polytomous mixed item model tests on the learning achievement test]. *Jurnal Penelitian dan Evaluasi Pendidikan*, 12(2), 302-320. <https://doi.org/10.21831/pep.v12i2.1433>
- Khiyarunnisa, A., & Retnawati, H. (2018). A two-tier diagnostic test instrument on calculus material: What, why, and how? In A. W. Subiantoro (Ed.), *5th ICRIEMS Proceedings* (pp. 479-485). Faculty of Mathematics and Natural Sciences. <https://bit.ly/3liFMS8>
- Khusnah, M. (2019). The development of two-tiers diagnostic test for identifying tenth-grade student's misconception about the categorization of hadith. [Master's thesis, Malang State Islamic University]. UIN Malang Digital Archive. <https://bit.ly/3sEcNMA>
- Lestari, S. A., Zawawi, I., Khikmiyah, F., & Fauziyah, N. (2021). Development evaluation tool two-tier multiple choice using wondershare quiz creator to identify mathematical connection. *Journal of Mathematics Education*, 6(2), 133-148. <https://bit.ly/3Psrc8s>
- Linacre, J. M. (2012). *Winstep: Rasch-model computer programs*. Winsteps.Com. <https://bit.ly/3wljq95>
- Malhotra, N. K. (2006). *Riset pemasaran* [Marketing research]. Erlangga.
- Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically* (2nd ed.). Pearson Educational Limited.
- Muhson, A., Lestari, B., Supriyanto, & Baroroh, K. (2017). The development of practical item analysis program for Indonesian. *International Journal of Instruction*, 10(2), 199-210. <https://doi.org/10.12973/iji.2017.10213a>
- Myanda, A. A., Riezky, M. P., & Maridi. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th high school. *International Journal of Science and Applied Science: Conference Series*, 4(1), 44-55. Sebelas Maret University. <https://bit.ly/3yIePzb>
- Oktaria. (2016). *Development of mathematics teaching materials with ICT for students in vocational schools* [Unpublished master's thesis]. University of Lampung.
- Permendikbud. (2016). *Standar isi pendidikan dasar dan menengah* [Standards of content for primary and secondary education]. Indonesian Government publicationservice. <https://bit.ly/3FZHlhp>
- Plomp, J. (2013). *Educational design research: An introduction*. Institute for Curriculum Development (SLO). <https://bit.ly/3MzqkNm>
- Putri, B. S., Kartono, & Supriyadi. (2020). Analysis of essay test instruments using higher order thinking skill (HOTS) at high school mathematics students using the Rasch model. *Journal of Research and Educational Research Evaluation*, 9(2), 58-69. <https://bit.ly/37QCtyu>
- Putri, E. L., Dwijanto, D., & Sugiman. (2017). Analysis of mathematical communication skills and confidence of 10th grade of SMK in geometry material viewed from cognitive style. *Unnes Journal of Mathematics Education*, 6(1), 97-107. <https://bit.ly/3yF4iF5>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya* [Item response theory and its application]. NuhaMedika. <https://bit.ly/39TFDIF>

- Rosidin, U. (2017). *Evaluasi dan asesmen pembelajaran* [Learning evaluation and assessment]. Media Akademi.
- Rovita, C. A., Zawawi, I., & Huda, S. (2020). Pengembangan alat evaluasi pembelajaran matematika berbasis two-tier multiple choice menggunakan Ispring Suite 9 [Development of an evaluation tool for learning mathematics based on two-tier multiple choice using Ispring Suite 9]. *Postulat: Jurnal Inovasi Pendidikan Matematika*, 1(2), 150-163. <https://doi.org/10.30587/postulat.v1i2.2094>
- Saepuzaman, D., Istiyono, E., Haryanto, Retnawati, H., & Yustiandi. (2021). Analisis estimasi kemampuan siswa menjawab soal fisik dengan pendekatan item response theory [Analysis of the estimation of students' ability to answer physical questions with an item response theory approach]. *Karst: Journal of Physics Education and Its Application/Karst: Jurnal Pendidikan Fisika dan Terapannya*, 4(1), 8-13. <https://doi.org/10.46918/karst.v4i1.948>
- Sarea, M. S. (2018). Karakteristik soal ujian akhir semester pendidikan agama islam dan budi pekerti tingkat sekolah dasar [Characteristics of the final exam for Islamic religious education and character at the elementary school level]. *An-Nahdhah*, 11(2), 303-318. <https://bit.ly/3FTouU7>
- Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Teori tes klasik and respon [Characteristics of items: Classical and response test theory]. *Didaktika: Jurnal Kependidikan*, 13(1), 1-16. <https://doi.org/10.30863/didaktika.v13i1.296>
- Suhaini, M., Ahmad, A., & Bohari, N. M. (2021). Assessments on vocational knowledge and skills: A content validity analysis. *European Journal of Educational Research*, 10(3), 1529-1540. <https://doi.org/10.12973/eu-jer.10.3.1529>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan* [Rasch modeling applications in educational assessment]. Trim Komunikata.
- Sundari, Kahar, M. S., & Erwinda, E. (2021). Analisis kemampuan berpikir tingkat tinggi menggunakan instrumen HOTS berbasis two-tier diagnostic test [Analysis of higher order thinking skills using the HOTS instrument based on a two-tier diagnostic test]. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, 10(4), 2726-2735. <https://doi.org/10.24127/ajpm.v10i4.4260>
- Suwarto. (2012). *Pengembangan tes diagnostik dalam pembelajaran* [Development of diagnostic test in learning]. Graha Ilmu.
- Syahlan. (2017). Sepuluh strategi dalam pemecahan masalah matematika [Ten strategies for solving math problems]. *Indonesian Digital Journal of Mathematics and Education*, 4(6), 358-369. <https://doi.org/10.31227/osf.io/6qfpm>
- Syaifuddin, M. (2020). Implementation of authentic assessment on mathematics teaching: Study on junior high school teachers. *European Journal of Educational Research*, 9(4), 1491-1502. <https://doi.org/10.12973/eu-jer.9.4.1491>
- Treagust, D. (1988). Development and use of diagnostic test to evaluate student's misconceptions in science. *International Journal of Science Education*, 10(2), 159-169. <https://doi.org/10.1080/0950069880100204>
- Untary, H., Risdianto, E., & Kusen. (2020). *Analisis data penelitian dengan model rash dan Winstep* [Analysis of research data with rash and Winstep models]. Halaman Moeka Publishing. <https://bit.ly/3NlnGLc>
- Vani, I. K., Paloloang, B., & Idris, M. (2019). Pengaruh persepsi dan minat belajar terhadap hasil belajar matematika siswa kelas X SMKN 6 Palu [The influence of perception and interest in learning on mathematics learning outcomes for students of class X Vocational School 6 Palu]. *Jurnal Elektronik Pendidikan Matematika Tadulako*, 6(4), 455-468. <https://bit.ly/3wjuXp9>
- Wulan, A. (2018). *Menggunakan asesmen kinerja untuk pembelajaran sains dan penelitian* [Using performance assessment for science learning and research]. UPI Press.
- Wulanningtyas, M. E., Suswanti, & Marhaeni, N. H. (2020). The student of error analysis and remedial program in working on the story of comparative material and turning value. *Jurnal Daya Matematis/Journal of Mathematical Power*, 8(2), 161-166. <https://doi.org/10.26858/jdm.v8i2.14344>
- Yang, T. C., Fu, H. T., Hwang, G. J., & Yang, S. J. (2017). Development of an interactive mathematics learning system based on a two-tier test diagnostic and guiding strategy. *Australasian Journal of Educational Technology*, 3(1), 60-80. <https://doi.org/10.14742/ajet.2154>

Appendix

Instrument of the Open Polytomous Response Test

Name of Student :

Class/Department :

School :

Instructions:Mark (x) one of the correct answer choices, and give the reason

(use another piece of paper to write down your reason)

- | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>1. Given an arithmetic sequence: 3, 8, 13, 18, ...The formula for the nth term of the sequence is</p> <p>A. $U_n = 5n - 3$ D. $U_n = 4n - 1$
 B. $U_n = 5n - 2$ E. $U_n = 3n + 2$
 C. $U_n = 4n - 1$</p> <p>Reason:</p> | <p>2. Given an arithmetic sequence: 2, 5, 8, 11, ...,68.The number of terms in the sequence is...</p> <p>A. 12 D.23
 B. 13 E.24
 C. 22</p> <p>Reason:</p> |
| <p>3. An arithmetic sequence, the 5th term is 31 and the 8th term is 46. The difference between the sequences is...</p> <p>A. 5 D.8
 B. 6 E.11
 C. 7</p> <p>Reason:</p> | <p>4. The 4th and 9th terms of an arithmetic sequence are 110 and 150. The 30th terms of the arithmetic sequence are...</p> <p>A. 308 D.344
 B. 318 E.354
 C. 326</p> <p>Reason:</p> |
| <p>5. An arithmetic sequence is: 3, 8, 13, 18, ..., 103.Then the middle term of the sequence is ...</p> <p>A. 53 D.11
 B. 52 E.10
 C. 20</p> <p>Reason:</p> | <p>6. Given the arithmetic sequence: 4, 10, 16, 22, ...If two numbers are inserted in every two consecutive terms then the 10th term of the sequence is...</p> <p>A. 18 D.24
 B. 20 E.26
 C. 22</p> <p>Reason:</p> |
| <p>7.The nth term of an arithmetic series is $U_n = 3n - 5$. The formula for the sum of the first n terms of the series is ...</p> <p>A. $S_n = \frac{n}{2}(3n - 7)$ D. $S_n = \frac{n}{2}(3n - 3)$
 B. $S_n = \frac{n}{2}(3n - 5)$ E. $S_n = \frac{n}{2}(3n - 3)$
 C. $S_n = \frac{n}{2}(3n - 4)$</p> <p>Reason:</p> | <p>8. The sum of the first n terms of an arithmetic series. $S_n = n^2 - 19n$. The formula for the nth term of the Series is.....</p> <p>A. $5n - 20$ D. $2n - 20$
 B. $5n - 10$ E. $2n - 10$
 C. $2n - 30$</p> <p>Reason:</p> |
| <p>9.The sum of all integers between 100 and 300 which are divisible by 5 is ...</p> <p>A. 8,200 D.7,600
 B. 8,000 E.7,400
 C. 7,800</p> <p>Reason:</p> | <p>10. PT. Angkasa Jaya in the first year produced 5,000 units. In the second year, production was reduced by 80 units per year. In what year did the company produce 3,000 units?</p> <p>A. 24 D.27
 B. 25 E.28
 C. 26</p> <p>Reason:</p> |
| <p>11. The middle term of an arithmetic sequence</p> | <p>12. A number of candies are distributed among</p> |

is 25.If the difference is 4 and the 5th term is 21.Then the sum of all the terms in the sequence is ...

- A. 175D.295
- B. 189E.375
- C. 275

Reason:

five children according to the rules of an arithmetic sequence. The younger the child, the more candy he gets. If the second child receives 11 pieces of candy and the fourth child 19 pieces, then the total number of candies is... pieces

- A. 60D.75
- B. 65E.80
- C. 70

Reason:

13. The sum of the first n terms of a series is $2n^2-n$.So the 12th term of the series is...

- A. 564D.45
- B. 276E.36
- C. 48

Reason:

14. The number of terms in thegeometric sequence:3, 6, 12, ..., 3072 is ...

- A. 9D.12
- B. 10E.13
- C. 11

Reason:

15. A geometric sequence has the 2nd term is 8, and the 5th term is 64. The 7th term of the sequence is ...

- A. 32D.256
- B. 64E.512
- C. 128

Reason:

16. The value of the middle term of the geometric sequence:

- 6, 3, ..., $\frac{3}{512}$ is ...
- A. $\frac{1}{16}$ D. $\frac{4}{16}$
 - B. $\frac{2}{16}$ E. $\frac{5}{16}$
 - C. $\frac{3}{16}$

Reason:

17. A piece of wire is cut with the first piece being 8 cm, and the next piece 1.5 times the previous cut. Then the 5th piece is... cm

- A. 18D.35
- B. 24E.40.5
- C. 27.5

Reason:

18. The sum of an infinite geometric series is 12 and the first term is 9. The ratio is ...

- A. $\frac{3}{4}$ D. $-\frac{1}{2}$
- B. $\frac{1}{4}$ E. $-\frac{3}{4}$
- C. $\frac{1}{3}$

Reason:

19. A ball falls from a height of 10 m and bounces back $\frac{3}{4}$ times its previous height. The total number of paths until the ball stops is... m

- A. 60D. 90
- B. 70E. 100
- C. 80

Reason:

20. Given $\begin{bmatrix} 4 & 8 & 5 \\ 6 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 2a & b+a & 5 \\ 6 & 3 & 1 \end{bmatrix}$.

The value of $3a + b$ is ...

- A. 8D. 14
- B. 10E. 20
- C. 12

Reason:

21. Given $K = \begin{bmatrix} a & 2 & 3 \\ 5 & 4 & b \\ 8 & 3c & 11 \end{bmatrix}$ and $L = \begin{bmatrix} 6 & 2 & 3 \\ 5 & 4 & 2a \\ 8 & 4b & 11 \end{bmatrix}$. If $K = L$, then cis ...

- A. 12D. 15
- B. 13E. 16
- C. 14

Reason:

22. Given $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 3 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 5 & 2 \\ -1 & 0 \end{bmatrix}$. Then $(A + C) - (A + B)$ is ...

- A. $\begin{bmatrix} 5 & 4 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix}$
- B. $\begin{bmatrix} 4 & 7 \\ 2 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 7 & -1 \\ 1 & -1 \end{bmatrix}$
- C. $\begin{bmatrix} 4 & 0 \\ -4 & -4 \end{bmatrix}$

Reason:

23. Given matrix $P = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$, then $(P^t)^t$ is...

- A. $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ D. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 1 & 3 & 1 \\ 3 & 3 & 7 \\ 1 & 5 & 6 \end{bmatrix}$ E. $\begin{bmatrix} 0 & 7 & 6 \\ 2 & 4 & 5 \\ 1 & 3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & 3 & 1 \\ 2 & 4 & 5 \\ 0 & 7 & 6 \end{bmatrix}$

Reason:

25. If $A = \begin{bmatrix} 3 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix}$, then $2AB$ is...

- A. $\begin{bmatrix} 13 & 42 \end{bmatrix}$ D. $\begin{bmatrix} 13 & 84 \end{bmatrix}$
 B. $\begin{bmatrix} 26 & 84 \end{bmatrix}$ E. $\begin{bmatrix} 30 & 36 \end{bmatrix}$
 C. $\begin{bmatrix} 26 & 42 \end{bmatrix}$

Reason:

27. Matrix X that satisfies $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} X = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$ is ...

- A. $\begin{bmatrix} -6 & -5 \\ 5 & 4 \end{bmatrix}$ D. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 B. $\begin{bmatrix} 5 & -6 \\ 4 & 5 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix}$
 C. $\begin{bmatrix} -6 & -5 \\ 4 & 5 \end{bmatrix}$

Reason:

29. Given matrix $A = \begin{bmatrix} 0 & 2 & 3 \\ -2 & 0 & 4 \\ -3 & -4 & 0 \end{bmatrix}$, then matrix determinant A is ...

- A. 0 D. 2
 B. 1 E. 4
 C. 2

Reason:

31. Given $A = \begin{bmatrix} 4 & 3 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 2 \\ 3 & 1 \end{bmatrix}$. Inverse matrix $(AB)^{-1} = \dots$

- A. $\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$ D. $-\frac{1}{2} \begin{bmatrix} 6 & 11 \\ 16 & 29 \end{bmatrix}$
 B. $\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$ E. $-\frac{1}{2} \begin{bmatrix} -6 & 11 \\ 16 & -29 \end{bmatrix}$
 C. $-\frac{1}{2} \begin{bmatrix} 29 & 11 \\ 16 & 6 \end{bmatrix}$

Reason:

33. The roots of the quadratic equation

24. Given matrix $A = \begin{bmatrix} 2x & -5 \\ 3 & y \end{bmatrix}$, $B = \begin{bmatrix} y & 2 \\ 2 & 4 \end{bmatrix}$, and $C = \begin{bmatrix} 8 & -3 \\ 5 & 2x \end{bmatrix}$.

If $A + B = C$, then $x + y = \dots$

- A. -5 D. 3
 B. -1 E. 5
 C. 1

Reason:

26. If $A = \begin{bmatrix} 1 & -3 \\ -2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -2 & 0 \\ 1 & 3 \end{bmatrix}$, and $C = \begin{bmatrix} 3 & -1 \\ 1 & -2 \end{bmatrix}$ then $A(B - C) = \dots$

- A. $\begin{bmatrix} -5 & -14 \\ 10 & 18 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -2 \\ -2 & 2 \end{bmatrix}$
 B. $\begin{bmatrix} -5 & -4 \\ 10 & 6 \end{bmatrix}$ E. $\begin{bmatrix} -7 & 19 \\ -2 & 20 \end{bmatrix}$
 C. $\begin{bmatrix} 1 & -16 \\ -2 & 22 \end{bmatrix}$

Reason:

28. If matrix $A = \begin{bmatrix} 3 & -x \\ 6 & 8 \end{bmatrix}$ is singular matrix. Value of x that satisfies is ...

- A. -5 D. 3
 B. -4 E. 4
 C. -3

Reason:

30. Transpose matrix P is P^T . If $P = \begin{bmatrix} 4 & 7 \\ 3 & 5 \end{bmatrix}$ then matrix $(P^T)^{-1}$ is ...

- A. $\begin{bmatrix} -5 & 7 \\ 3 & -2 \end{bmatrix}$ D. $\begin{bmatrix} -3 & 5 \\ 4 & -7 \end{bmatrix}$
 B. $\begin{bmatrix} 3 & -4 \\ -5 & 7 \end{bmatrix}$ E. $\begin{bmatrix} 4 & -3 \\ -7 & 5 \end{bmatrix}$
 C. $\begin{bmatrix} -5 & 3 \\ 7 & -4 \end{bmatrix}$

Reason:

32. The roots of the quadratic equation $3x^2 - 4x + 3 = 0$ are ...

- A. $x^2 + x - 12 = 0$
 B. $x^2 - x - 12 = 0$
 C. $x^2 - x + 12 = 0$
 D. $x^2 - 3x + 4 = 0$
 E. $x^2 - 4x + 3 = 0$

Reason:

34. The roots of the equation $x^2 - 2x - 3 = 0$

$5x^2 + 4x - 12 = 0$ are ...

- A. -2 and $\frac{5}{6}$
- B. 2 and $-\frac{5}{6}$
- C. 2 and $\frac{6}{5}$
- D. -2 and $-\frac{6}{5}$
- E. -2 and $\frac{6}{5}$

Reason:

are x_1 and x_2 . If $x_1 > x_2$, then $x_1 - x_2$ is ...

- A. -4
- B. -2
- C. 0
- D. 2
- E. 4

Reason:

35. The roots of the quadratic equation: $2x^2 - 3x - 9 = 0$ are x_1 and x_2 . Value of $x_1^2 + x_2^2$ is ...

- A. $11\frac{1}{4}$
- B. $6\frac{3}{4}$
- C. $2\frac{1}{4}$
- D. $-6\frac{3}{4}$
- E. $-11\frac{1}{4}$

Reason:

36. The roots of the quadratic equation: $x^2 + 2x + 3 = 0$ are α and β . The quadratic equation $(\alpha - 2)$ and $(\beta - 2)$ is ...

- A. $x^2 + 6x + 5 = 0$
- B. $x^2 + 6x + 7 = 0$
- C. $x^2 + 6x + 11 = 0$
- D. $x^2 - 2x + 3 = 0$
- E. $x^2 + 2x + 11 = 0$

Reason:

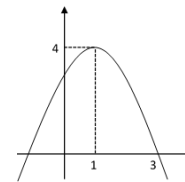
37. A quadratic function that has a minimum value 2 for $x = 1$ and has a value of 3 for $x = 2$ is ...

- A. $y = x^2 - 2x + 1$
- B. $y = x^2 - 2x + 3$
- C. $y = x^2 + 2x - 1$
- D. $y = x^2 + 2x + 1$
- E. $y = x^2 + 2x + 3$

Reason:

38. The figure below is a graph of the quadratic equation? ...

- A. $y = x^2 + 2x + 3$
- B. $y = x^2 - 2x - 3$
- C. $y = -x^2 + 2x - 3$
- D. $y = -x^2 - 2x + 3$
- E. $y = -x^2 + 2x + 3$



Reason:

39. If f is a quadratic function whose graph passes through the points (1,0), (4,0) and (0, -4) then the value of $f(7)$ is ...

- A. -16
- B. -17
- C. -18
- D. -19
- E. -20

Reason:

40. The graph equation of a quadratic function has an extreme point (-1, 4) and through (0, 3) is ...

- A. $y = -x^2 + 2x - 3$
- B. $y = -x^2 + 2x + 3$
- C. $y = -x^2 - 2x + 3$
- D. $y = -x^2 - 2x - 5$
- E. $y = -x^2 - 2x + 5$

Reason:

SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Sun, May 22, 2022 at 4:08 PM

To: Editor - European Journal of Educational Research <editor@eu-jer.com>

Dear Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research

I have checked and no more errors were found, so this article is ready for publication.
Thank you for your kindness.

Best regards,

Sugeng Sutiarso
University of Lampung
[Quoted text hidden]



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Online First of the Manuscript EU-JER ID#21112502244011

1 message

European Journal of Educational Research <editor@eu-jer.com>
Reply-To: European Journal of Educational Research <editor@eu-jer.com>
To: European Journal of Educational Research <sugeng.sutiarso@fkip.unila.ac.id>
Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Mon, May 23, 2022 at 1:24 AM

Dear Dr. Sugeng Sutiarso,

We have published your finalized paper entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" (Manuscript EU-JER ID#21112502244011) as online first in our web site (See <https://www.eu-jer.com/volume-11-issue-3-july-2022>).

We have assigned the doi number for your paper. This link also belongs to your paper's web page:
<https://doi.org/10.12973/eu-jer.11.3.1441>

Please find the pdf of your paper at this link:
https://eu-jer.com/EU-JER_11_3_1441.pdf

We will publish officially the new issue on July 15, 2022.

Please check your web page and paper carefully. If your paper/ web site has any mistake, please do not hesitate to contact me immediately.

Best regards,

Ahmet C. Savas, Ph.D.
Editor, European Journal of Educational Research
editor@eu-jer.com
www.eu-jer.com



SUGENG SUTJARSO <sugeng.sutiarso@fkip.unila.ac.id>

Congratulations! Your paper has been published (EU-JER ID#21112502244011)

1 message

European Journal of Educational Research <editor@eu-jer.com>

Fri, Jul 15, 2022 at 11:43 PM

Reply-To: European Journal of Educational Research <editor@eu-jer.com>

To: European Journal of Educational Research <sugeng.sutiarso@fkip.unila.ac.id>

Cc: undang.rosidin@fkip.unila.ac.id, aansulistiawan95@guru.smp.belajar.id

Dear Dr. Sugeng Sutiarso,

Congratulations! We have published your paper entitled "Development of Mathematics Assessment Instruments for Learning with Polytomous Response in Vocational School" (Manuscript EU-JER ID#21112502244011) on our website officially (See <https://www.eu-jer.com/volume-11-issue-3-july-2022>).

We have published the valuable articles from 23 different countries (Australia, China, Cyprus, Greece, Indonesia, Iraq, Israel, Japan, Jordan, Kingdom of Saudi Arabia, Kosovo, Malaysia, Mexico, Pakistan, Romania, Russia, Slovenia, South Korea, Spain, Ukraine, United Arab Emirates, USA, and Vietnam).

We have assigned the doi number for your paper. This link also belongs to your paper's web page:

<https://doi.org/10.12973/eu-jer.11.3.1441>

Please find the pdf of your paper at this link:

https://eu-jer.com/EU-JER_11_3_1441.pdf

Could you publicize our journals to your colleagues please? We are looking forward to getting your contributions to EU-JER in the future.

Best regards,

Ahmet C. Savas, Ph.D.

Editor, European Journal of Educational Research

editor@eu-jer.com

www.eu-jer.com