AKSARA *Jurnal Bahasa dan Sastra*
Vol. 20, No. 2, pp. 72 – 87, October 2019
http://jurnal.fkip.unila.ac.id/index.php/aksara

Reflection of the Test-Item Quality in State SMP and SMA in Bandar Lampung

Ujang Suparman
Lampung University, Bandar Lampung, Indonesia
ujang.suparman@fkip.unila.ac.id

Abstract. The objectives of this research are to analyze critically the quality of test items used in SMP and SMA (mid semester, final semester, and National Examination Practice) in terms of reliability as a whole, level of difficulty, discriminating power, the quality of answer keys and distractors. The methods used to analyze the test items are item analysis (ITEMAN), two types of descriptive statistics for analyzing test items and another for analyzing the options. The findings of the research are very far from what is believed, that is, the quality of majority of test items as well as key answers and distractors are unsatisfactory. Based the results of the analysis, conclusions are drawn and recommendations are put forward.

Key words: assessment, quality of test items, test item analysis

INTRODUCTION

This research is focused on analyzing the quality of multiple choice (MC) test items or objective test items (OTI). Although some experts may put forward their objections on the use of MC test items (Hinchliffe 2014, Srivastava et al., 2004), the MC tests are still widely used in an examination (Rodriguez, 2005; Mehta, et al. 2014; Kaur, et al. 2016; Namdeo, et al. 2016; Rauch, et al. 2010; McKenna, 2019) or test which involves many participants, such as, final-semester examination at schools, school-final examination, university entrance test, and new employee recruitment. Therefore, this research is still relevant and up to date to meet the need of a good quality of test items. A good quality of English instruction should be accompanied with good quality of assessment (Wiliam, 2013, He, et al. 2018; Black, et al. 1998; Quaigrain, et al. 2017). A good quality of assessment can be seen from high index in validity, reliability, (Bolarinwa, 2015; Mohajan, 2017; Bajpai, et al. 2014; Taherdoost, 2016) level of difficulty and level of discriminating power (Boopa-thiraj, et al. 2013; Khoshaim, et al. 2016; Chauhan, et al. 2013). Besides, a good quality of assessment is indicated by good quality of key answers and distracters (Hasan, et al. 2017; Chauhan, P., et al. 2015; Rahma, et al. 2017; Burud, et al. 2019; Rao, et al. 2016; D'Sar, et al. 2017). Assessment cannot be separated from instruction because assessment is intended to measure whether the intended outcomes of the instruction are achieved or not.

In other words, if the assessment is in line with the instruction, and meet the necessary quality of a good and effective assessment, the results of the assessment must reflect the objectives of the instruction perfectly. By contrast, if the assessment is not in congruent with the instruction and it does not meet the required quality for effective and optimal assessment, it does not and will never reflect the intended outcome of the instruction.

Many studies have been conducted on the quality of assessment (Çanakkale, et al. 2013; Büyükkarcı, 2014; Patnaik, et al. 2015; Ibili, et al 2019; Gokdas, et al. 2019; Haidari, et al. 2019, Astawa, et al. 2017), but few studies, if any, have been conducted pertaining to the quality of test items, key answers, and distractors used for mid semester, and final semester as well as national examination for SMP and SMA. The current study tried to deal with the unresolved issues.

Büyükkarcı (2014) investigates teachers' beliefs on assessment o student achievement, he has found that although assessment has a primary role in education and cannot be separated from instruction activities, teachers of language do not apply principles of assessment as required by the currcilum. The condition of student assessment perhaps may not be different between what happens in Turkey and in Indonesian context where language teachers remain unaware of the importance of the quality of test item quality in English education. Patnaik, et al. (2015) carry out a study from teacher perspective. They investigate the parameters of teacher quality and put priority on the necessity of updating oneself regularly to meet the challenges of teaching profession including the teachers' mastery on student chievement assessment. Ibili, et al (2019) have conducted a study of the relationship of feeling of ease and cognitive load of students. They have found that there is a strong correlation between the feeling of ease of test items and the extraneous load in males, and there is a strong relationship between the feeling of usefulness and the intrinsic load in females. Both the feeling of usefulness and the feeling of ease of use of test items have a strong correlation with the student cognitive mastery of language. This means that level of difficulty and discriminating power of a test item is very important for students to solve English test materials.

In line with the review of the literature above, it was assumed that there was a discrepancy between the theories of the assessment with the reality in the field. Particularly in relation to the quality of a stem of a test, that is, it must be proportional – neither too difficult nor too easy. If the test item is too difficult, it may not be answerable by majority of the participants including the clever ones. By contrast, if it is too easy, it may be answered correctly by both clever and non-clever students. In other words, if a test item is too difficult or too easy, it may result in its poor discriminating power, because it cannot be used to discriminate between the clever and non-clever students. Or it may happen that the resulted index is negative (-), that is, when an individual or group of clever students cannot answer an item correctly, but an individual or group of non-clever students can answer correctly. Such a case suggests that the test item does not have a good discriminating power.

Besides, there are two other components of a multiple choice test item which are almost neglected by an English teacher when designing and administering an objective test, that is, options comprising of the key answer and distractors (Burud, et al. 2019; Rahma, et al. 2017). Based on an informal focused group discussion (FGD) in the field, test designers are not aware of the important role of the options (source: an informal FGD with SMP and SMA English teachers in an MGMP meeting). They may have thought that the most important component of a test item is the key answer. Consequently they do not focus seriously on constructing good quality of distractors. Based on the theory, all the options should function well, which is indicated by being chosen by at least 5% of the testees. If any option is merely chosen by less than 5% of the testees, or even no one chooses it, it suggests that they may have thought that the answer is not that one. In short, it is very clear for them the inappropriateness of the distractors; consequently they do not choose it. Although this issue is theoretically very important for assessing the student achievement, no special research, at least which has been ever published, focuses on the issue. Therefore, this research dealt, among others, with the unresolved issue.

Theoretically, Gronlund et al. (2009: 93-106) put forward rules for designing multiple choice (MC) items: 1. Construct a test item to assess a significant learning achievement; 2. Put forward only one clearly formulated problem in the stem of the item; 3. Express the stem of the item in an easily understandable language; 4. Use as much of the wording as possible in the stem of the item, avoid repeating the same material in each of the choices; 5. If possible, state the in stem of the item in an affirmative form; 6. When negative wording is used in the stem of an item, it should be emphasized; 7. Make sure that the key answer is correct or clearly best; 8. All options should be grammatically correct and in line with the stem of the item and similar in form; 9. Prevent from using verbal clues that may cause the students to select the correct answer or to eliminate the incorrect options; 10. Make the distracters interesting for the uninformed; 11. Differentiate the relative length of the correct answer to remove length as the clue; 12. Prevent from using “all of the above” and use “none of the above” with great attention; 13. Change the position of the correct answer in a random manner; 14. Control the difficulty of item either by changing the problem is the stem or by varying the options; 15. Make sure that each item is independent of the other items in the test; 16. Apply an efficient item format; and 17. Use normal grammatical rules. There may be some other rules which are not included in the list, but this is enough for general guidelines.

The objectives of this study were, first, to analyze the reliability of the test items as a whole, then the quality of each of the test items, in terms of level of difficulty, level of discriminating power; after that, the quality of the answer key, and finally the quality of the distracters. After each of these objectives was identified, then it was followed by decisions or recommendations.

METHODOLOGY

This research used a descriptive and evaluative method, that is, a study which described the results of an evaluation on a certain object which was adjusted with standard criteria. The objects of the current research were English-test items consisting of one unit of mid-semester exam for SMPN, one unit of final semester exam for SMPN, one unit of mid-semester exam for SMAN, one unit of final semester exam for SMAN, and one unit of National Exam Practice (LUN). These five different units of English test item model were intended to identify whether the quality of each unit similar to or different from one another. And finally to predict what may happen in the future if the results of the analysis of such test items were interpreted. The outcome of the current research is expected to support the theory of assessment in general and to be a beneficial feedback for curriculum developers and test-item designers in practical.

This research used a documentary procedure, that is, five different units of test items and students’ answers in their answer sheets from five different groups depending on the types of test items relevant to the levels of participants. That is, students’ answer sheets for SMPN mid semester exam, for SMPN final semester exam, for SMPN national exam practice (LUN), for SMAN mid semester exam, and for SMAN final semester exam. The data pertaining to the quality of the stems of the test items were analyzed using item analysis (*Iteman*) software, called Micro Computer Adaptive Test (MicroCat) version 3.50A, and interpreted using standard criteria of assessment. *Iteman* itself can be defined as one of “the analysis programs that comprise assessment systems of test items and test analysis package,” (Assessment Systems Corporation (ASC) (1989-2006).

The reliability of each of the test unit was analyzed using the Iteman software, the results of which were compared with the standard criteria in Table 1.

Table 1. Criteria for determining reliability

Alpha (test item reliability) → Decision	
0.000 – 0.400	Low/not sufficient
0.401 – 0.700	Average/sufficient
0.701 – 1.000	High/Good

The discriminating power of each of the test items was analyzed using the Iteman software version 3.50A. Then the results of the statistical calculation were consulted with the following standard criteria in Table 2.

Table 2. Criteria for determining discriminating power

Point Biserial (Discriminating Power – D) → Decision	
Parameter of D	→ Decision
– 0.199	Very low /needs dropping or total revising
0.200 – 0.299	Low/needs revising
0.300 – 0.399	Quite average/without revision
0.400	High /very good

Like analyzing the discriminating power, to analyze the level of difficulty, the Iteman software version 3.50A was used. And then to determine the decision, the results of the statistical computation were consulted with the standard criteria in Table 3 below:

Table 3. Criteria for determining Level of Difficulty

Prop Correct (Level of Difficulty – p)	
Parameter of p	→ Decision
0.000 – 0.099	Very difficult/needs total revising
0.100 – 0.299	Difficult/needs revising
0.300 – 0.700	Average/good
0.701 – 0.900	Easy/needs revising
0.901 – 1.000	Very easy/ needs dropping or total revising

Finally, to analyze the quality of distractors, the Iteman software version 3.50A was used. After that, to determine the decision, the results of the statistical computation were consulted with the standard criteria below:

Table 4. Criteria for determining the quality of distracters

Prop Endorsing (proportion of the answers)	
Parameter of p	→ Decision
0.000 – 0.010	Least/drop, or needs revising
0.011 – 0.050	Sufficient/good enough
0.051 – 1.000	Very Good

RESULTS AND DISCUSSION

As stated in the background of the research, there are five objectives of the current research, using five different units of test items, the results of the data analysis and discussion are organized in a similar systematic way.

First, results of SMPN mid-semester exam data analysis

1. There were 34 examinees in the data file. From the scale of the statistics, it can be inferred that the score of alpha is 0.727, it means that the reliability of the test items is high/good. It suggests that the test items as a whole are good and can be used but some items which are problematic should be revised first.
2. There are 9 items out of 50 (18%) that are considered good and can be used directly without any revision and can be put on the exercise bank.
3. 24 items out of 50 (48%) should be revised first before being used because one of the prop correct (level of difficulties)and point biserial (level of discriminating power) of the items cannot achieve good criteria, (see Tables 3 and 4).
4. 17 items out of 50 (34%) should be dropped because they do not fulfill the criteria of level of difficulty and the level of discriminating power, (see Tables 3 and 4).
5. There are 38 key answers out of 50 (76%) which are considered good and can be directly used without any revision; 12 out of 50 key answers (24%) are poor because they do not have good discriminating power; 37 distractors out of 200 (18.5%) work well and therefore, can be directly used without any revision; and 163 distractors out of 200 (81.5%) do not work well because there are some distractors that have the prop endorsing and point biserial indexes of 0.00 (very low). Which means that the distracters were attractive for the testees or they felt sure that they were obviously wrong.

These results of the analysis show that the number of good quality test items is less than that of those that should be revised and dropped. This also implies that the teachers' mastery of good and effective assessment should be developed. In other words, the principles of good and effective assessment have not been applied. Besides, the quality of key answers is also necessary to be re-trained to the teachers and prospective teachers. It has been found that not all key answers are good, that is, there are 24% of them are still poor because their discriminating power is low, that is, they cannot discriminate between the clever and non-clever students. In other words, it is quite possible that both groups cannot answer correct or both of them can answer correctly. It is interesting to note that 81.5% of distracters do not work well. It means that majority of the testees do not choose them which may be due to the clarity of the wrong choice that makes them not to choose it.

Second, results of SMPN final semester exam data analysis

1. There were 32 examinees. From the scale of the statistics, it can be inferred that the score of alpha is 0.427, it means that the reliability of the test items is average. It suggests that the test items as a whole are good and can be used but some items which are problematic should be revised first.
2. There is 1 item out of 25 (4%) that is considered good and can be used directly without any revision and can be put on the exercise bank.
3. 6 out of 25 items (24%) should be revised because one of the prop correct and point biserial of the items cannot achieve good criteria, (see Tables 3 and 4).
4. There are 18 items out of 25 (72%) that should be dropped because the items do not fulfill the criteria of prop correct and point biserial, (see Tables 3 and 4).
5. There are 9 key answers out of 50 (18%) which are considered good and which can be directly used without any revision; 41 out of 50 key answers (82%) are poor because they do not have good discriminating power; 9 out of 100 distracters (9%) are good because they are chosen by at least 5% of the participants; 16 distractors out of 100 (16%) should be revised because the point biserial indexes belong to low category; 75 distractors out of 100 (75%) should be dropped because there are some distractors that have the prop endorsing and point biserial indexes of 0.00 (very low).

These results make us more surprised because there is only one item out 25 items which is categorized good. The rest are poor. And there are 18 items (72%) that should be dropped because they are too poor. This suggests that the test items are not tried out first before they are administered. The teacher(s) designed the test items and then they directly administered to assess their students' achievement. When the key answers are compared between the good and the poor ones, the poor (82%) surpasses the good (18%). This is a real challenge for the LPPTK to re-consider the English Teaching Assessment subject. It means that this topic should be more focused so that teachers are aware of the importance of the key answers and distracters. The distracters should be conceptually, and grammatically correct so that the students who are well prepared will choose it. Thus, the distracters function well, (Rahma, et al. 2017; Chauhan, P., et al. 2015; Burud, et al. 2019; Rao, et al. 2016; D'Sar, et al. 2017).

The following pie diagrams show the proportions of the quality of the test items used for final semester exam in SMPN.

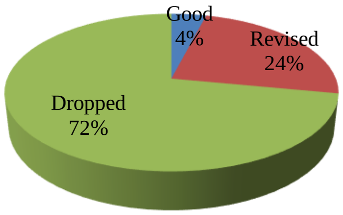

Figure 1a. Analysis of the test items

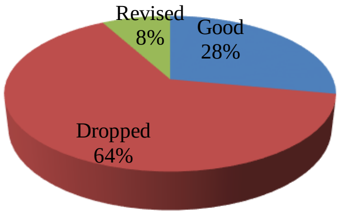

Figure 1b. Analysis of the key answers

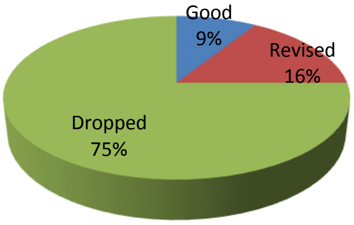

Figure 1c. Analysis of the distractors

The following are the results of SMPN LUN data analysis

1. There were 36 examinees in the data file. From the scale of the statistics we can conclude that the score of alpha is 0.274, which means that the level of the test items is low (not sufficient).
2. There are 5 items out of 50 (10%) that are considered good and can be used directly without any revision and can be put on the test-item bank.
3. 10 items out of 50 (20%) that should be revised because one of the prop correct and point biserial of the items cannot achieve good criteria.
4. 35 items out of 50 (70%) that should be dropped because they do not fulfill the criteria of prop correct and point biserial.

5. There are 35 key answers out of 50 (70%) which are considered good and can be directly used without any revision; 15 out of 50 key answers (30%) are poor because they do not have good discriminating power; 40 distractors out of 200 (20%) work well and which can be directly used without any revision; and 160 distractors out of 200 (80%) which do not work well because there are some distractors that have the prop endorsing and point biserial indexes of 0.00 (very low). Therefore, they should be revised or changed with new ones. The following are the representations of the SMPN LUN:

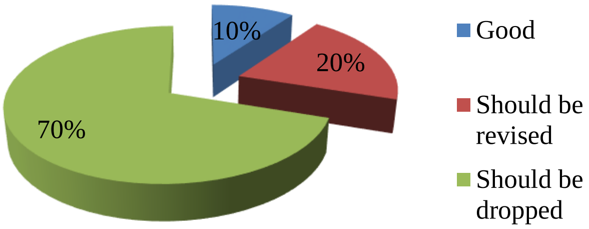

Figure 2a. Result of Data Analysis – LUN SMPN Test Items

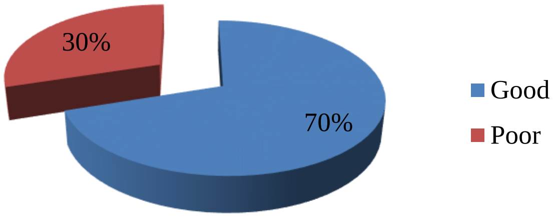

Figure 2b. Result of Data Analysis – LUN SMPN Key Answers

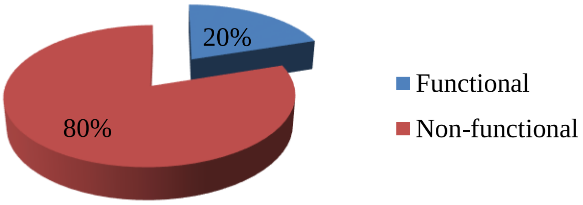

Figure 2c. Results of Data Analysis – LUN SMPN Distracters

Similar to the results of the analysis of test items for SMPN, the following is the results of SMAN Mid Semester Exam.

1. Based on the results of data analysis, it was found that there were 31 examinees in the data file. From the scale of the statistics, it can be stated that

the score of Alpha (reliability) is 0.544, it means that the level of the test items is Average.

2. Besides the reliability, it was found that there were 13 out of 50 items (26%) which were considered good and can be used directly without any prior revision in terms of level of difficulty (Prop. Correct) and discriminating power (Point Biser).
3. 21 items out of 50 items (42%) should be revised first before being used, because most of their point biserials are very low or needs revising.
4. 16 out of 50 items (32%) should be dropped because their point biserials are less than 0.200. Therefore, those items do not fulfill the criteria of test item's quality and should be dropped.
5. 14 key answers out of 50 (28%) are considered good, therefore, can be directly used without any revision; 6 key answers out of 50 (12%) should be revised because the point biserial indexes belong to low; and 28 key answers out of 50 (56%) should be dropped because the point biserial indexes belong to very low. Besides, there are 52 distractors out of 200 (26%) which belong to good category, therefore they can be directly used without any revision; 148 distracters out of 50 (74%) should be dropped because their prop endorsing and point biserial indexes are 0.00 (very low). It means no one chosed them. The following are the figures showing the results of Mid semester of SMAN:

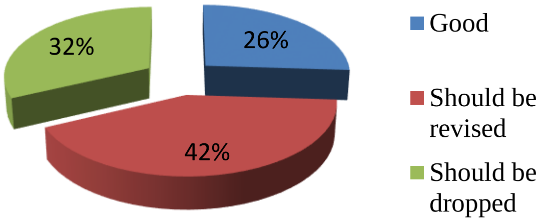

Figure 3a. Results of Data Analysis – MID SMAN Test Items

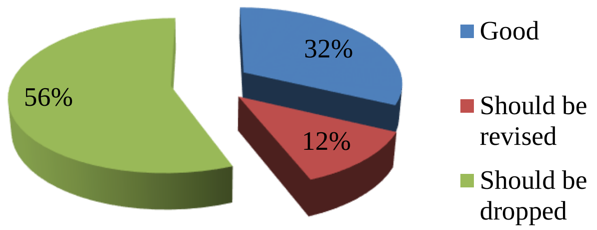

Figure 3b. Results of Data Analysis – MID SMAN Key Answers

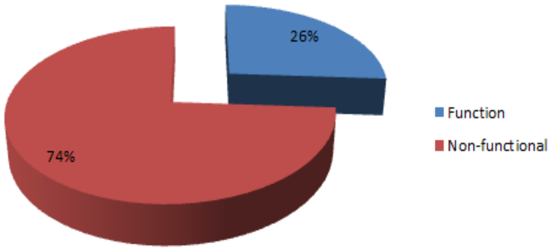

Figure 3c. Results of Data Analysis – MID SMAN Distracters

Finally, the results of the last analysis – SMAN Final Semester Exam (UAS) item test analysis – found the following points:

1. The alpha index of the whole test items (reliability) is 0.799 which belongs to high or good.
2. There were 27 items out of 50 items (54%) which were considered good and can be used directly without any prior revision in terms of level of difficulty (prop. Correct) and discriminating power (point biserial).
3. 18 items out of 50 items (36%) should be revised first before being used, because most of their point biserial is very low or needs revising.
4. 5 items out of 50 (10%) should be dropped because their point bisers are less than 0.200. Therefore, those items do not fulfill the criteria of test item's quality and should be dropped.
5. There are 23 out of 50 key answers (46%) belong to good category and therefore can be used directly without any revision; 20 out of 50 key answers (40%) should be revised because they do not have sufficient discriminating power 7 items (14%) whose key answers should be dropped and changed with new ones because there was a message **Please check the answer keys.* Concerning with the distracters, there are 92 out of 200 distracters (46%) which work well; 12 out of 200 distracters (6%) do not work well and should be revised; and 56 out of 200 distracters (28%) should be dropped since they were non-functional. It means they were not selected at all by all examinees.

The results of the analysis of the test items for SMAN final semester examination are relatively better than the rest of the test items used. This can be seen from the number of test items that can be used directly without any revision, 54%, and those which should be revised are 36% while those items that should be dropped are 10%.

Based on these results of the data analysis of the five different units of test items above, it can be interpreted that the concepts of a good quality test items need to be shared with the English teachers in almost all schools in every level of education. The teachers should be trained intensively to have a good mastery of analyzing any test items (daily test, mid semester test and final semester or school examination test items). If they do not have sufficient ability to analyze the test items, they could not measure precisely the intended learning outcome.

Based on these findings, there are some discrepancies between the theories of assessment and the realities in the field. This is a challenging task for Teacher Training and Education Faculty and other LPTK (institutions whose responsibility is to produce high quality teachers for all school levels from kinder garden through general senior and vocational high schools), for curriculum designers, and policy makers as well as other stake holders who are dealing with education.

CONCLUSIONS AND SUGGESTIONS

In line with the objectives of this research stated in the background section, that is, to analyze the reliability of the test items as a whole, then the quality of each of the test items, in terms of level of difficulty, level of discriminating power; after that, the quality of the answer key, and finally the quality of the distracters, which are then followed by decisions, and because there were five different units of test items used, (SMPN Mid Semester Exam, SMPN Final Semester Exam, SMPN National Examination Practice (LUN), SMAN Mid Semester Exam, and SMAN Final Semester Exam), the following conclusions are drawn:

1. The construct of the test items, to some extent, does not contain many mistakes not only on the stems but also on the options. Consequently, it tends to be applicable based on the construct of the test items, provided that the test items which contain some mistakes should be revised and re-tried out to make sure the optimal quality before they are administer-ed.
2. Special attention should be made priority on making sure that reliability, level of difficulty, discriminating power, key answers and distractors.
3. Based on the findings, it can also be interpreted that, to a certain extent, the stems and options of the test items which consisted of key answers and distracters are still away from the theories of good quality assessment.
4. Categorically speaking, the quality of the test items based on the results of the analysis can be categorized as follows:
   a. Good test items which can be directly used without prior revision (SMPN Mid Semester Exam: 18%; SMPN final semester exam: 4%; SMPN LUN: 10%; SMAN Mid Semester Exam: 26%; SMAN Final Semester Exam: 54%).

b. Good test items, but should be revised first and re-tried out before being used (SMPN Mid Semester Exam: 48%; SMPN Final Semester Exam: 24%; SMPN LUN: 20%; SMAN Mid Semester Exam: 54%; SMAN Final Semester Exam: 36%).
c. Bad test items – cannot be used and must be dropped because of too bad level of difficulties, too bad discriminating power (SMPN mid semester exam: 34%; SMPN Final Semester Exam: 72%; SMPN LUN: 26%; SMAN Mid Semester Exam: 32%; SMAN final semester exam: 10%).
d. Good key answers (SMPN Mid Semester Exam: 76%; SMPN Final Semester Exam: 18%; SMPN LUN: 70%; SMAN Mid Semester Exam: 28%; SMAN Final Semester Exam: 10%).
e. Bad key answers and therefore must be dropped and changed with new ones (SMPN Mid Semester Exam: 24%; SMPN Final Semester Exam: 82%; SMPN LUN: 30%; SMAN Mid Semester Exam: 72%; SMAN Final Semester Exam: 30% should be revised and 14% should be dropped).
f. Good distracters which can be directly used without prior revision (SMPN mid semester exam: 18.5%; SMPN Final Semester Exam: 9%; SMPN LUN: 20%; SMAN Mid Semester Exam: 26%; SMAN final semester exam: 46%).
g. Poor distracters which can be directly used without prior revision (SMPN mid semester exam: 81.5%; SMPN Final Semester Exam: 16% should be revised and 75% should be dropped; SMPN LUN: 80%; SMAN Mid Semester Exam: 8% should be revised and 66% should be dropped; SMAN final semester exam: 12% should be revised and 22% should be dropped).

RECOMMENDATIONS

1. Given that this research only focuses on test items of English assessment for SMPN and SMAN, further reserachers are recommended to investigate those used for vocational schools and religious-institution based schools (e.g. M.Ts. and MAN).
2. Further reserachers are also recommended to carry out research of authentic assessment since this research only focuses on multiple choice assessment.
3. Given that the findings show that in all different five units of test items, only less than half of the items are good, the rests are poor even should be dropped, likewise the quality of the options are more than fifty percent poor even should be dropped, anaysis of test items should be made more socialized to teachers and prospective teachers by LPTK (teacher training and education institutions) and by education authority in province and regency levels.

REFERENCES

Astawa, I. N., Handayani, N. D., Mantra, I. B. N., & Wardana, I. K. (2017). Writing English language test items as a learning device: A principle of habit formation rules. *International Journal of Social Sciences and Humanities, 1*(3), pp. 135-144. https://doi.org/10. 29332 /ijssh.v1n3.67

Bajpai, S. & Bajpai, R. (2014). Goodness of measurement: Reliability and validity. *International Journal of Medical Science and Public Health, Volume 3, Issue 2, 22014*, pp. 112-115. DOI: 10.5455/ijmsph. 2013.191120133

Black, P. & Wiliam, D. (1998). Assessment and classroom Learning. *Assessment in Education, Vol. 5, No. 1, 1998*, pp. 7-73. Journal homepage: http://www. Tandfon-line.com/loi/caie20

Bolarinwa, O.K. (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. Nigerian Postgraduate Medical Journal, Vol. 22, Issue 4, 2015. pp. 195-201. DOI: 10.4103/ 1117-1936.173959

Boopathiraj, C. & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research, (*IRJC) *Vol.2 (2), February (2013),* pp. 189-193. Online available at indianresearch journals.com

Burud, I., Nagandla, K. & Agarwal, P. (2019). Impact of distractors in item analysis of multiple choice questions. *International Journal of Research in Medical Sciences, 2019 Apr;7(4)*, pp. 1136-1139. www.msjonline.org DOI: http://dx.doi.org/10.18203/2320-6012. Ijrms 20191313

Büyükkarcı, K. (2014). Assessment beliefs and practices of language teachers in primary education. *International Journal of Instruction, January 2014, Vol.7, No.1*, pp. 107-120. http://*www.e-iji.net*

Çanakkale, G.T. & Çanakkale, G.M. (2013). Developing a science process skills test regarding the 6th graders. *The International Journal of Assessment and Evaluation, Volume 19, 2013, pp. 39-57*. http://thelearner.com/

Chauhan, P.R, & Bhoomika, C. (2013). Study of difficulty level and discriminating index of stem type multiple choice questions of anatomy in Rajkot. *Biomirror, Volume 4(06), pp. 1-4(2013)*, pp. 1-4.

Chauhan, P., Chauhan, G.R., Chauhan, B.R., Vaza, J.V. & Rathod, S.P. (2015). Relationship between difficulty index and distracter effectiveness in single best-answer stem type multiple choice questions. *International Journal of Anatomy and Research, Int J Anat Res 2015, Vol 3(4), pp. 1607-10*. DOI: http://dx.doi.org/10.16965/ijar.2015. 299

D'Sar, J.L., & Visbal-Dionaldo, M.L. (2017). Analysis of multiple choice questions: Item difficulty, discrimination index and distractor efficiency. *Internation-al Journal of Nursing Education, July-September 2017, Vol.9, No. 3*, pp. 109-114.

Gronlund, N.E. & Waug, C.K. (2009). *Assessment of student achievement.* Upper Saddle River, New Jersey: Pearson.

Hinchliffe, J. (2014), *CQ university scraps multiple choice exams in an Australian first*. available at: www.abc.net.au/ news/2014-09-23/cqu-scraps-multiple-choice-exams-in-an-australian-first/5763226 (accessed 28 July 2017), ABC News, 23 September 2014.

Ibili, E. & Billinghurst, M. (2019). Assessing the relationship between cognitive load and the usability of a mobile augmented reality tutorial system: A study of gender effects. *International Journal of Assessment Tools in Education, 2019, Vol. 6, No. 3*, pp. 378–395. https://dx.doi.org/ 10.214 49/ijate.594749; http://www. ijate. net; http:// dergipark.org.tr/ijate

Gokdas, I. & Kuzucu, Y. (2019.) Social network addiction scale: The validity and reliability study of adolescent and adult form. *International Journal of Assess-ment Tools in Education, 2019, Vol. 6, No. 3*, pp. 396–414. https://dx.doi.org/ 10.21449/ijate.505863 http://dergi park.org.tr/ijate

Haidari, S.M. & Karakuş, F. (2019). Safe learning environment perception scale (SLEPS): A validity and reliability study. *International Journal of Assessment Tools in Education, 2019, Vol. 6, No. 3*, pp. 444–460. https://dx.doi.org /10.21 449/ijate.505863; http://dergipark.org. tr/ijate

Hassan, S. & Hod, R. (2017). Use of item analysis to improve the quality of single best answer multiple choice question in summative assessment of undergraduate medical students in malaysia. *Education in Medicine Journal. 2017; 9(3)*, pp. 33–43. *www.eduimed.com* Penerbit Uni-versiti Sains Malaysia. 2017. https: //doi.org/10.21315/eimj 2017. 9.3.4

He, J., Barrera-Pedemonte, F. & Buchholz, J. 2018. Cross-cultural comparabil-ity of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice*, pp. 1-17. DOI: 10.1080/0969594X.2018. 1469467; Journal homepage: http:// www.tandfonline.com/loi/caie20

Kaur, M., Singla, S. & Mahajan, R. (2016). Item analysis of in use multiple choice questions in pharmacology. *International journal of Applied Basic Medical Research*, 2016, Vol. 6, Issue 3, pp. 170-173 Available from: http: // www.ijabmr.org/text.asp?2016/6/3/ 170/186965

Khoshaim, H.B. & Rashid, S. (2016). Assessment of the assessment tool: Analysis of items in a non-MCQ mathematics exam. *International Journal of Instruc-tion, January 2016, Vol.9, No.1, pp. 120-132 www.e-iji.net.* DOI:10.12973/iji.201 6.9110a

McKenna, P. (2019) "Multiple choice questions: Answering correctly and knowing the answer", *Interactive Technology and Smart Education*, https://doi.org/10.1108/ITSE-09-2018-0071; https://doi.org/10.1108/ITSE 09-2018-0071

Mehta, G. & Mokhasi, V. (2014). Item analysis of multiple choice questions-An assessment of the assessment tool. *International Journal of Health Sciences and Research, Vol. 4, Issue 7, 2014*, pp. 197-202. www. ijhsr.org

Mohajan, H. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University, 17(3), July 2017)*, pp. 58-82. https://mpra.ub.uni-muenchen.de/ 83458/

Namdeo, SK., & Sahoo, B. (2016). Item analysis of multiple choice questions from an assessment of medical students in Bhubaneswar, India. *International Journal of Research in Medical Sciences, 2016. 4(5),* pp1716-1719. DOI: http://dx.doi.org/10.18203 /2320-6012.ijrms20161256;

Patnaik, D.S. & Davidson, L.M. (2015). The role of professional development in ensuring teacher quality. *International Journal of English Language Teaching Vol.3, No.5, July 2015*. pp.13-19. www.eajour nals.org

Quaigrain, K. & Arhin, A.K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evalua tion. *Cogent Education* (2017), 4: 1301013, pp. 1-11. http://dx.doi.org/10.1080/2331186X.2017.1301013

Rahma, A., Shamad, M., Idris, M. E. A., Elfaki, O., Elfakey, W., & Salih, K. M. A. (2017). Comparison in the quality of distractors in three and four options type of multiple choice questions. *Advances in Medical Education and Practice, Volume 8, 287–291*. Doi:10.2147 /amep.s128318

Rao, C., Prasad, K. H. L., Sajitha, K., Permi, H. & Shetty, J. (2016). Item analysis of multiple choice questions: Assessing an assessment tool in medical students. *Interna-tional Journal of Educational and Psycho-logical Researches, Vol. 2, Issue 4, October-December 2016*, pp. 201-204. http://www.ijeprjournal.org

Rauch, D.P. & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, Volume 52, 2010 (4)*, pp. 354-379

Rodriguez, M.C. (2005). Three options are optimal for multiple-choice items: a meta-analysis of 80 years of Research. *Educational Measure-ment: Issues and Practice*, pp. 1-13. https:// doi.org/10.1111/j.1745-3992.2005. 00006.x

Srivastava, A., Dhar, A. and Aggarwal, C.S. (2004), “Why MCQ”, *Indian Journal of Surgery, Vol. 66,* pp. 246-248.

Taherdoost, H. (2016). Validity and reliability of the research instrument: How to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management (IJARM), Vol. 5, No. 3, 2016,* pp. 28-36. www. elvedit.com

Wiliam, D. (2013). Assessment: The bridge between teaching and learning. *Voices from the Middle, Volume 21 Number 2, December 2013,* pp. 15-20.