ANALYZING THE QUALITY OF ENGLISH TEST ITEMS OF DAILY, MID SEMESTER AND FINAL SCHOOL EXAMINATIONS IN BANDAR LAMPUNG: (ASSESSMENT AND EVALUATION IN LANGUAGE TEACHING)

By:
Ujang Suparman
Lampung University, INDONESIA

Abstract

The problem of the current study is that the test items used to measure the objectives of the teaching and learning in the English (such as what happens in schools in Lampung Province), are rarely pre-determined before they are used, and the results of the test are also rarely analyzed systematically and professionally. Besides, few eachers of English are aware of the importance of good quality of test items such as validity, reliability, discriminating power and level of difficulty. Even some teachers consider that analyzing such matters is time consuming and hard to do. Consequently, they rarely analyze either the test items or the results of the test. The objective of the research is to potray some examples of quality of the test items after being analyzed using *Iteman*. The current presentation is based on a huge study using the *iteman* in English test items used in Elementary, Junior High and Senior High Schools in Lampung. It has been found that in general the test items has high validity, sufficient relaibility shown by the Alpha in each test used for each level, Standar Deviation, average mean of each item, and Mean Biserial (average mean of the whole test items). Besides, it has also been found that some items can be used directly without any revision; some need revision before being ussed; and even some others need dropping due to the quality of the test items. It is recomended that teachers, and instructors who test their students using multiple choice items should consider the quality of test items using one of standard technques such as the *iteman software* as an effective alternative for item test analysis because it is easy, fast, and comprehensive analysis to do.

Key words: *item quality, option quality, validity, reliability, descriminating power*

INTRODUCTION

Evaluation plays a major role in education. It shows whether the objectives of the teaching and learning can be achieved or not. And in its turn, evaluation can be used to improve students' performance. Therefore, the instrument used should meet the criteria of a good test item (Suparman, 2013). However, the teachers are sometimes unaware of the importance of the quality of the test they use, consequently, some of them rarely test the instruments before they use them (see Suparman, 2013).

This research covers the analysis of huge and diverse English test items used in elemntary, junior high and senior high schools. There are 45 university English Educaion Program students participating in data collection and data analysis held in 2015 under the guidance of a lecturer of English Teaching Assessment. It involved 15 schools and 1,800 students. All the test items are used for either daily test (prepared by the English teacher), mid semester exam, final semester exam, or final school exam (UAS) usually prepared by MKKS. The objective of the research was to identify the quality of test items and to develop them based on the information obtained from the results of the analysis using the *Iteman*. The analysis covers the major issues closely related to the assessment: validity, reliability, discriminating power and level of difficulty of test items; besides it also

includes the analysis of all the options comprising of the key answer to the question, and distractors. The current presentation is only a part of the total research but which can reflect and represent the figure of the quality of English test items used in Lampung.

THEORETICAL FOUNDATION

Many factors may influence the results of a test. For example, Kheirzadeh, et al (2015) state that the condition of a test administration, that is, the timing of the test, the testing venues and the exam proctors/inspectors are influential factors that may affect construct-irrelevant variance to a test, if ignored, and therefore resulted a test invalid.

WHAT IS *ITEMAN*?

According to Assessment Systems Corporation (ASC) (1989-2006), *iteman* can be defined as one of the analysis programs that comprises Assessment Systems Corporation's Item and Test Analysis Package. It is very important for lecturers and teachers of English who are responsible for administering tests (such as mid semester and final semester examinations) to know what *iteman* is; why it is important; how it works, and what the example of an item analysis using *iteman*. Basically, *iteman* can be used to analyze tests and survey item response data and provide conventional item analysis statistics (e.g., proportion/percentage endorsing and item-total correlations) for each item. Such function is very important for English teachers at school levels in order to assist them in determining the extent to which items are contributing to the reliability of a test and which response alternatives are functioning well for each item. Besides item-level statistics, more importantly the *iteman* program also provides statistical indicators on the performance of the test as a whole (e.g., mean, standard deviation, reliability, median p-value).

The Procedure of Using *Iteman*

The data that have been gathered in order to be analyzable by *iteman* should be formatted in a special file called ASCII (text-only) files. This can be accomplished perfectly by using a Notepad, an *iteman* for Windows text editor, that is, a word-processing editor that provides true ASCII output, or a program written specifically to format your data. It is also highly necessary to note that all the data that would be analyzed must be contained in a single input file. One of good points of it is that a single analysis can cover up to 750 items, while the number of examinees is almost unlimited.

A data file in an *iteman* can be put under five primary components:
1. A control line describing the data;
2. A line of keyed responses;
3. A line of the numbers of alternatives for the items;
4. A line specifying which items are to be included in the analysis; and
5. The examinee data, (ASC, 1989-2006: 2).

An example of a data file on an *iteman*

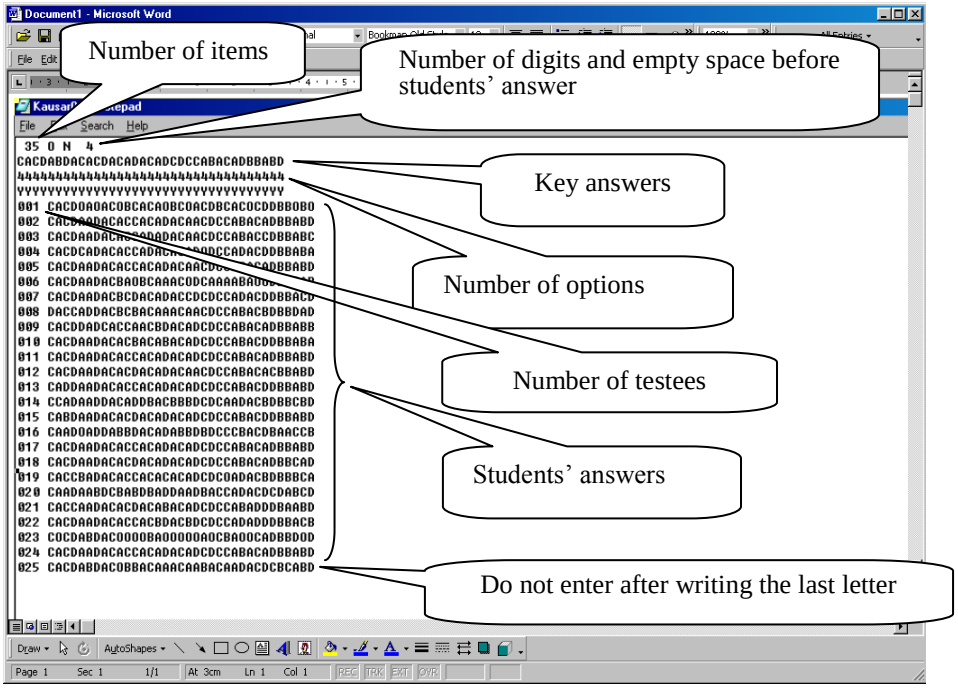


Fig 1. Data file using Notepad on Windows
Source: Results of data analysis

STEPS OF ANALYSIS WITH AN *ITEMAN*

The *iteman* program can work only with multiple choice items. It is relatively easy to analyze test items using the *iteman* program. The most important thing to do is to be very careful in entering the data, because if you enter the data wronghly, it would produce wrong results of data analysis. The following are the steps to enter the data using a new file:

1. Click *Start*
2. Select *Program*
3. Select *Accessories*
4. Choose and click *Notepad*
5. Save/ Click *File*
6. Select and click *Save as*, then name the data file, for example: Advread (make sure the file name must not exceed 8 letters/numbers)
7. Startt data entry, it will be faster if you work with your friend – one of you reads students' answers and the other types them. If you work with your friend, please make sure to prounuce the letter clearly, e.g., a for apple; b for ball; c for charlie; d for doctor; and e for ent.
8. It's advisable for you to save it frequently by clicking *File* and then *Save* so that the typed data will not loss if the current suddenly cuts off.
9. The data will appear like shown on the Fig. 1 above.

Procedure of Data Analysis Using *Iteman* Program

The steps used in the current study are as follows:

1. Open *iteman* Program, by clicking *Start*,
2. Select *Program*/click *iteman*.
3. Type the name of your data file (input) as you like on *Enter the name of the input file*. For example D:\English.txt then *Enter*.
4. Enter the name of the output file on *Enter the name of the output file*. For example, inthis case: D:\English.output then click *Enter*.
5. A question will appear, *Do you want the scores written to a file? (Y / N).* Then type *Y* and click *Enter*.
6. Enter the name of your score file on *Enter the name of the score file:* For example, D:\English.scr
   Then click Enter. Finish. Have a good try!

   The data would appear like that in the following page.

MicroCat (tm) Testing System
Copyright ©1982,1984, 1986, 1988 by Assessment Systems Corporation
Beta-Test version – Univ. of Pittsburgh
Item and Test Analysis Program -- ITEMAN (tm) Version 3.00

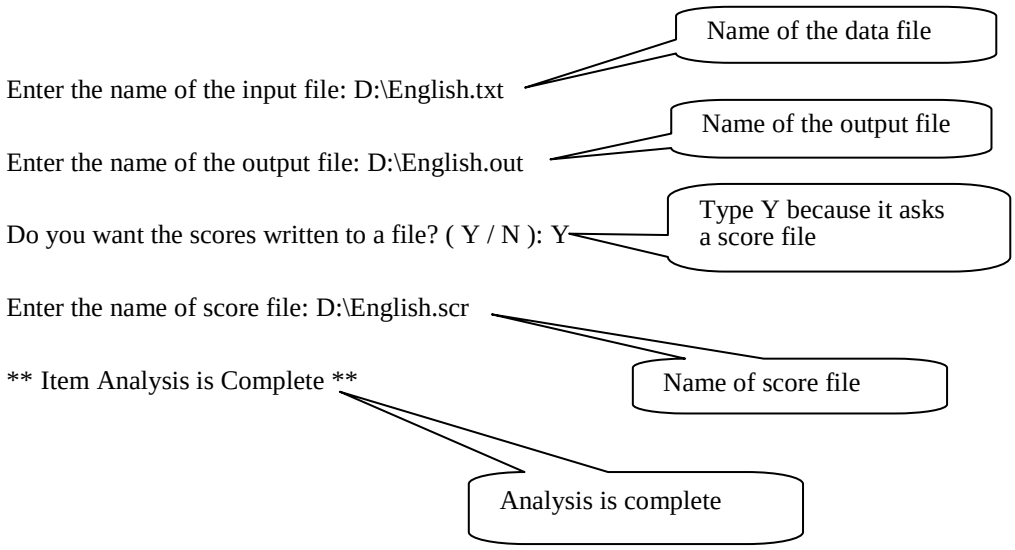


Fig. 2. Item analysis appearance using ITEMAN

The Results of Item Analysis with *Iteman*

The following is the steps that the researcher used to open the results of item analysis on MS Words program:

1. Click *Start*,
2. Select *Program/*click Microsoft Word
3. Click *File/*click *Open*, please look for the results on, for example, Drive D (depends on which one you choose).
4. The following is an example of the appearance of the results of test items analysis.

MicroCAT (tm) Testing System Page 1
Copyright (c) 1982, 1984, 1986, 1988, 1993 by Assessment Systems Corporation
Item and Test Analysis Program -- ITEMAN (tm) Version 3.50
Item analysis for data from file F:\Advraed.TXT
Date: 10/22/12 Time: 8:35 am

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.		Prop. Total	Endorsing Low	Point High	Biser.	Key
1	0-1	.32	.35	.36	A	.16	.30	.00	-.31	
					B	.32	.20	.55	.36	*
					C	.03	.00	.09	.12	
					D	.49	.50	.36	-.15	
					Other	.00	.00	.00		
2	0-2	.19	.17	.11	A	.05	.00	.09	.19	?
					B	.70	.80	.55	-.27	
	CHECK THE KEY				C	.19	.10	.27	.11	*
	C was specified, A works better				D	.05	.10	.09	.16	
					Other	.00	.00	.00		
4	0-4	.27	.45	.26	A	.22	.30	.27	.10	
					B	.03	.00	.00	-.02	
					C	.27	.00	.45	.26	*
					D	.49	.70	.27	-.31	
					Other	.00	.00	.00		
5	0-5	.78	.10	.19	A	.78	.90	1.00	.19	*
					B	.00	.00	.00		
					C	.16	.10	.00	-.17	
					D	.05	.00	.00	-.07	
					Other	.00	.00	.00		
6	0-6	.68	.41	.30	A	.08	.10	.09	.12	
					B	.05	.00	.00	-.01	
					C	.68	.50	.91	.30	*
					D	.19	.40	.00	-.44	
					Other	.00	.00	.00		

And so on.

In the following, the resume of the results of the item tests analysis is presented and at the right side are the scores obtained by each of the participants.

Scale:	:	0
N of Items	:	60
N of Examinees	:	37
Mean	:	37.703
Variance	:	51.506
Std. Dev.	:	7.177
Skew	:	-0.677
Minimum	:	19
Maximum	:	51
Median	:	38
Alpha	:	0.798
SEM	:	3.224
Mean P	:	0.628
Mean Item-Tot.	:	0.288
Mean Biserial	:	0.405
Max Score (Low)	:	34
N (Low Group)	:	10
Min Score (High)	:	41
N (High Group)	:	11

No	Scores	Marks
1	37	62
2	24	40
3	45	75
4	40	67
5	20	33
6	33	55
7	45	75
8	19	32
9	41	68
10	31	52

Students' scores

Table 1. The resume of the results of the item tests analysis

More importantly, ITEMAN can function as a powerful technique available to teachers for improving the quality of instruction. To achieve this, the items that would be analyzed should fulfill the following requirements: first, they have to be valid measures of instructional objectives; secondly, they have to be diagnostic, in the sense that, knowledge of which incorrect options that the students choose must be a clue to the nature of the misunderstanding, and, therefore, prescriptive of appropriate remediation; and finally, teachers who construct their own examinations may greatly improve the effectiveness of test items and the validity of test scores if they select and rewrite their items on the basis of item performance data.

VALIDITY

One of the characteristics of a good test is validity. It requires a test to be usable to measure what it is intended to measure and nothing else (Power, 2012). In a similar concept, Hatch, et al (1982:250-1) define validity as "the extent to which the results of the procedure serve the uses for which they were intended." They further divide the validity into three basic types: content validity, criterion-related validity and construct validity. *Content validity* can be defined as the extent to which a test measures a representative sample of the subject matter content, that is, the test items should be relevant to the materials covered in the course. For example, in the case of the reading comprehension, the test should include the sub-skills of reading as stated in the syllabus, among others: identifying the main idea, identifying specific information relating to *who, what, when, why, where* and *how* questions; making predictions, and making inferences. The second type of validity is *criterion-related* validity. It is defined as the criteria of a test when test scores will be used to predict future performance or to estimate current performance on some valued measure other than the test itself. For example, we have designed a new language aptitude test for the students of English Study Program. And the test is thought to be a good one. Then the test is administered to a group of newly enrolled students at

the English Program, and to proove that it is a valid test, the results are compared with an established test, say English Language Placement Test, which is the criterion expected to be able to predict. We predict from our aptitude test scores to performance on the major subjects at the English Program.

The two types of validity above enable us to determine how well test scores represent certain learning objectives (content validity) or how well they predict or estimate a certain performance (criterion-related validity). Besides these more specific and practical uses, sometimes the validity of 'certain general psychological construct' (Hatch, et al, 1982: 252) needs to be identified. For example, when the students' performances in terms of psychological aspects (such as *self-esteem, extrovert/introvert, acculturated, motivated*) need to be interpreted, and how important they are in language learning in English classes), then construct validity is required. But this type of validity is not the concern of the current study.

RELIABILITY

A clear cut and direct definition of reliability is "consistency of measurement" (Su, et al, 2015), In this study the definition of *reliability* is straightforward: a measurement is reliable if it represents mostly true score, relative to the error. For example, an item such as "Red foreign cars are particularly ugly" would likely provide an unreliable measurement of prejudices against foreign-made cars. This is because there probably are many individual differences concerning the likes and dislikes of colors. Thus, this item would "capture" not only a person's prejudice but also his or her color preference. Therefore, the proportion of true score (for prejudice) in subjects' response to that item would be relatively small.

At least, *Reliability & Item Analysis* have three major functions. First, they may be used to construct reliable measurement scales, secondly, to improve existing scales, and finally to evaluate the reliability of scales already in use. In a more specific objective, *Reliability & Item Analysis* will aid in the design and evaluation of *sum scales*, that is, scales that are made up of multiple individual measurements (e.g., different items, repeated measurements, and different measurement devices). Numerous statistics can be computed to allow us to build and evaluate scales following the so-called *classical testing theory* model.

MEASURES OF RELIABILITY

Based on the discussion above, one can easily infer a measure or statistic to describe the reliability of an item or scale. Specifically, an *index of reliability* may be defined in terms of the proportion of true score variability that is captured across subjects or test takers, relative to the total observed variability. In equation form, it can be stated:

Reliability = $\sigma^2_{(\text{true score})}$ / $\sigma^2_{(\text{total observed}).}$

but this equation is not used in this study because that is automatically calculated by the *iteman* software.

NUMBER OF ITEMS AND RELIABILITY

This concept describes a basic principle of test design. That is, the more items there are in a scale designed to measure a particular concept, the more reliable will the measurement be. Let us examine the following example to clarify the concept. Suppose you want to

measure the height of 10 persons, using only a crude stick as the measurement device. Note that we are not interested in this example in the absolute correctness of measurement (i.e., in inches or centimeters), but rather in the ability to distinguish reliably between the 10 individuals in terms of their height. If each person is measured only once in terms of multiples of lengths of your crude measurement stick, the resultant measurement may not be very reliable. However, if each person is measured 100 times, and then take the average of those 100 measurements as the summary of the respective person's height, then you will be able to make very precise and reliable distinctions between people (based solely on the crude measurement stick).

DISCRIMINATING POWER

There are two indicators of the item's discrimination effectiveness: point biserial correlation and biserial correlation coefficient (Matlock-Hetze, 1997). The choice of correlation is determined by what kind of question we want to answer. The advantage of using discrimination coefficients over the discrimination index (D) is that every person taking the test is used to compute the discrimination coefficients and only 54% (27% upper + 27% lower) are used to compute the discrimination index, D.

The point biserial (rpbis) correlation is used to find out if the right people are getting the items right, and how much predictive power the item has and how it would contribute to predictions. The discriminating power (D) of test items can be measured by one of the three ways: discriminating index; correlation index; and harmonious index. A discriminating power is usually symbolized with a capital D, which can be determined by the following steps: First, rank order the answer sheet top-down from the highest to the lowest scores based on the total number of test takers; then multiply N with 27%, the results is n score; after that, calculate n from the Upper Group (the answer sheets with high scores are counted from the top) while n from the Lower Group (the answer sheets with low scores are counted from the bottom). And finally, determine the proportion of the test items answered correctly by each group. That is, the correct answers from each of the Upper Group (pU) and Lower Group (pL) are divided by n. the discriminating power is in fact the differences of the proportion of the correct answers between the UG and the LG. So, it can be stated that D = pU – pL.

To determine whether a test item is accepted, revised or rejected, the following parametric criteria is used:

Parameter of D Coefficient	**Decision**
D = > 0.30	accepted
D = 0.10 - 0.29	revised
D = <0.10	rejected

LEVEL OF DIFFICULTY

Level of item difficulty can be defined as the percentage of students taking the test who answered the item correctly. In short, in can be stated that the larger the percentage getting an item right, the easier the item. The higher the difficulty index, the easier the item is understood to be. Matlock-Hetzel (1997) states that to compute the item difficulty, the examiner can divide the number of people answering the item correctly by the total

number of people answering item. The proportion for the item is usually denoted as p and is called item difficulty (Crocker & Algina, 1986). An item answered correctly by 85% of the examinees would have an item difficulty, or p value, of .85, whereas an item answered correctly by 50% of the examinees would have a lower item difficulty, or p value, of .50.

The easiest way to measure the level of difficulty of an item is by using proportional scale or proportion correct (p), that is, the number of test takers answering correctly on the items under analysis is compared with the total number of test takers. The equation is as follows:

$$p = \frac{\sum B}{N}$$

where p = the proportion of test takers who answer correctly a certain item under analysis
$\sum B$ = the number of test takers who answer correctly
N = the total number of test takers.

The level of difficulty ranges from 0 through 1. It can be categorized into three classifications as follows:

Proportion Correct (p)		Category
$p \geq 0.70$	:	easy
$0.30 < p < 0.70$	:	Average
$p < 0.30$	:	difficult

METHOD

The design of the research is descriptive assessment, that is, a study describing the results of an analysis of the topic under discussion, which was adjusted with standardized criteria. The research analyzed the test items used to assess the students' ability to response to daily, mid semester, final semester and final school examinations in elemntary, junior high, and senior high schools. The tests were prepared by the teachers of English for daily, mid semester and final semester examinations. Whereas the tests for final school examination (UAS) are usually prepared by the board of headmasters (MKKS). The research took place in the schools mostly in Bandar Lampung; but some schools are out of Bandar Lampung. It was carried out during the First Semester of the 2015/2016 academic year. The researcher collaborates with 45 students from the English Study Program, University of Lampung taking English Teaching Assessment course. The total participants of the research consist of 1,800 students learning English classes in 15 schools, in each level (elementary, junior high and senior high schools).

The data were collected by means of documentation, that is, using the students' answer sheets on the English tests or examination comprising daily English tests, mid semester English test, final semester examination and final school examination on the academic year mentioned above. The data, that is, the students' answers and scores on English tests were analyzed using *Iteman* software. The analysis covered four major issues relating to the assessment: validity, reliability, discriminating power and level of difficulty.

HOW TO INTERPRET THE RESULTS OF ITEM ANALYSIS

Based on the recommendations from some experts of measurement, the following criteria to determine the quality of test items and its interpretation have been agreed:
to classify which test items can be used directly without prior revision, which ones need revising or even which ones need dropping.

Level of Difficulty (p)	
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
Point Biserial (Discriminating Power – D)	
0.199 –	Very low $\leq$ D/needs dropping or total revising
0.200 – 0.299	Low/needs revising
0.300 – 0.399	Quite average/without revision
0.400	High $\geq$ D/very good
Prop Endorsing (proportion of the answers)	
0.000 – 0.010	Least/drop, or needs revising
0.011 – 0.050	Sufficient/good enough
0.051 – 1.000	Very Good
Alpha (reliability of test item)	
0.000 – 0.400	Low/not sufficient
0.401 – 0.700	Average/sufficient
0.701 – 1.000	High/Good

Table 2. Criteria to classify the quality of test items

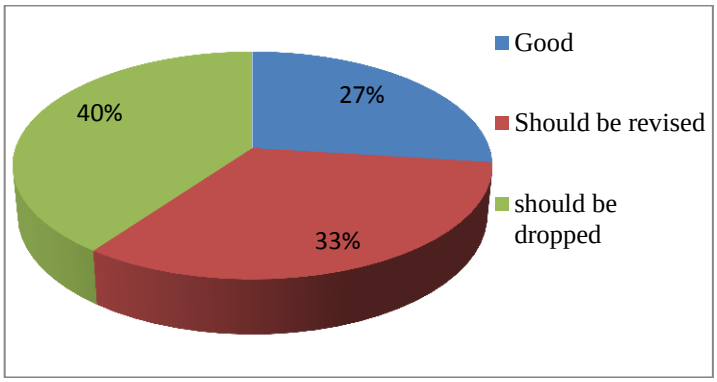
FINDINGS AND CONCLUSIONS

Based on the results of the data analysis on daily test, mid semester, final semester, and final school examinations, the following are found:

Example of Daily Test Results of Analysis

In one of elementary school, the results of test items analysis show that out of 30 English test items 8 items (good, not necessarily revised before being used), 9 items should be revised because the quality is not sufficient for a good test and 13 items should be dropped because the quality is too bad. The following figure shows the quality of the daily test items.

Fig. 1 Quality of Daily Test Items in an Elementary School in Lampung

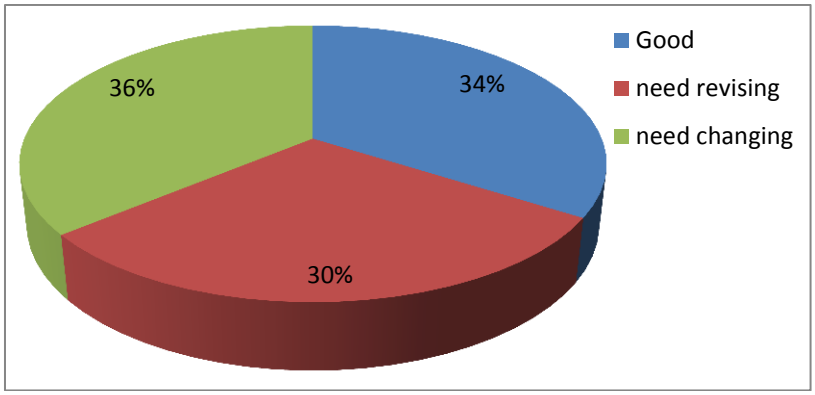


The figure above shows that good test items are relatively smaller in number than the items that should be revised. And what makes things worse is that the items that should dropped have the highest percentage. It is a challenge for teachers, school supervisors and Diknas to solve the problem.

Example of Mid Semester Exam Results of Analysis

In one of the state junior high schools, it was found that in a mid semester exam, there were 17 out of 50 test items (34%) which can be used directly without any revision.

Fig 2. Quality of Mid Semester Exam Items in an Elementary School in Lampung

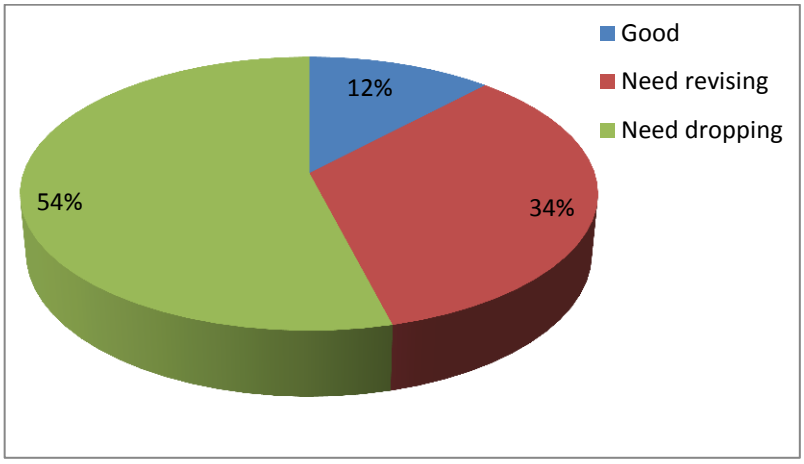


Besides, there are 15 out 50 English test items (30%) which should be revised before they are used in the examination. And unexpectedly, there are 18 out of 50 test items (36%) which should be dropped because their quality were very bad. There are several reasons for the problem including: The range between Prop Correct and Point Biserial is too far; the values of Point Biserial are undetected by the iteman because it is too small; there are negative points (minus) in several Point Biserial values; the values of Prop Correct is very high, even can reach 1.00. It means that the item is very easy. Both of the values of Prop Correct and Point Biserial are very low. There are some 0 values found in Point Biserial, which means that there is no different range in answering/choosing option. Low students and hig students all choose the right answer.

Example of Final Semester Exam Results of Analysis

The results of the data analysis in a final semester examination in one of senior high schools show that 6 out of 50 test items (12%) are good and therefore can be used directly without any revision; 17 out of 50 (34%) need revising because they are not good enough based on the predetermined criteria; and 27 out 50 (54%) should be dropped because the quality of the items are very bad. This is represented by the following figure.

Fig. 3 Quality of Final Semester Exam Test Items in a Senior High School in Lampung

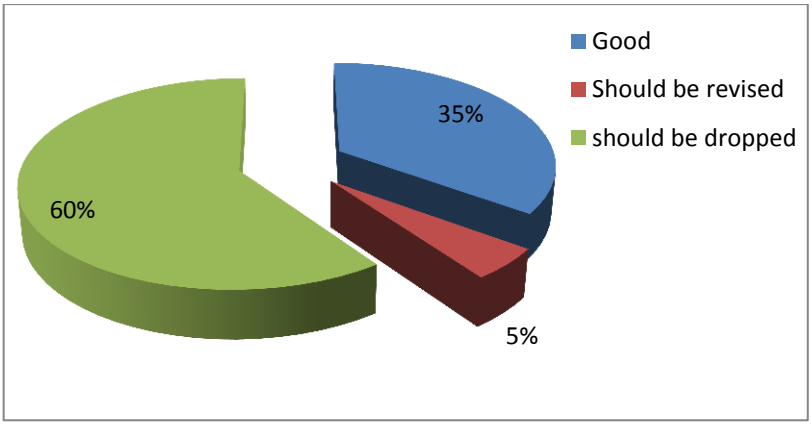


The figure also shows that unexpectedly 54% of the total number of the test items should be dropped which means that majority of the English test items used in a final semester examination are not good.

Example of Final School Exam Results of Analysis

Another result of data analysis of English test items used in a final school examination (UAS) has also shown similar findings to what has been discussed so far as illustrated by the following figure.

Fig. 4 Quality of Final School Examination Test Items in a Senior High School in Lampung



Like what has been explained so far, the last figure above also shows that majority of the English test items (60%) should be dropped because the quality of them are very bad.

Based on the findings above the following conclusions can be drawn and recommendations can be put forward:

1. Although the analysis of the English test items was carried in different levels (elementary, junior high and senior high schools), the main stream is similar, that is, the quality of English test items absolutely needs improving.
2. In all levels of educational institutions, the English test items can be categorized into three classifications: sound, that is, they can be used directly without prior revision, unfortunately the percentage is relatively low; need revising before being used (unexpectedly the percentage is very high); and need dropping because they are too bad (unfortunately the percentage is sometimes very high).
3. Given the teachers are very busy, they seem hardly try out the test items that they will use. They tend to design the tests and directly use them.
4. There is no enforcement from the school supervisors and/or Diknas for teachers to try out the test items, analyze the results and make sure the quality of the items.
5. Although evaluation belongs to one of the 8 standards of national education, in reality its implementation in the lowest layer – school – seems, to some extent, neglected.
6. It is recommended that teachers of English in all levels of educational institutions should always try out the tests that they have prepared before being used to measure the objectives of their teaching.
7. There should be special training for English teachers on how to analyze test items so that they can make sure that the tests that they design are sound – having high quality based on the standard of a good test, and whether their teaching objectives can be achieved.
8. The school supervisors and/or Diknas should put priority on the improving the quality of test items so that the quality of teaching and learning process can be improved.
9. School headmasters should make sure regularly at least once in one semester that the teachers have done the try out for any kind of test that they will use.
10. To improve English teachers' motivation, there should be working appraisal for those who have done the tryout of the tests and analyzed the results.

BIBLIOGRAPHY

ASC. 1989-2006. *User's Manual for the ITEMAN™ Conventional Item Analysis Program*. St. Paul, Minnesota: Assessment Systems Corporation

Crocker, L., and Algina, J. 1986. *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Hatch, E. & Farhady, H. 1982. *Research design and statistics for applied linguistics.* Rowley, Messachussetts: Newbury House Publishers.

Kheirzadeh, S., Marandi, S.S., & Tavakoli, M. 2015. Test Administration Conditions of the General English Section of the Iranian National PhD Entrance Exam: Are the PhD Exam Candidates Satisfied?. *Journal of Language Testing* Vol. 5, No. 2, October 2015

Lyman, H.B. 1971. *Test scores and what they mean*. Englewood Cliffs, New Jersey: Prentice Hall, Inc.

Matlock-Hetzel, S. 1997. *Basic Concepts in Item and Test Analysis*. Texas: A&M University

Miriam, K.L. 1996. *An introduction to psychological tests and scales*. London: UCL Press.

Ngadimun, H. 2004. *Analisis butir soal dengan komputer dan menafsirkannya.*Makalah disampaikan pada sosialisasi KBK bagi guru SMP Kabupaten Tanggamus di Pulau Panggung, tanggal 22-24 Juli 2004. Bandar Lampung: HEPI.

Power, T. 2012. *Methods of assessment*. http://www.tedpower.co.uk/esl0706.html (Accessed: 09 May 2013)

Su, Y & Shin, S.Y. 2015. Test Review: The New HSK. *Iranian Journal of Language Testing.* Vol. 5, No. 2, October 2015

Suparman, U. 2011. The implementation of Iteman to improve the quality of English test items as a foreign language. *Aksara*, Vol. XII, No. 1 pp. 85-96. Bandar Lampung, April 2011.

Suparman, U. 2013. *Improving the quality of English test items as a foreign language by means of Iteman: (an assessment analysis)*. A paper presented on the First International Teacher Education Conference (ITEC) organized by Lampung Univeristy in Bukit Randu Hotel, June 30th – July 3rd, 2013.