REiD (Research and Evaluation in Education), 4(1), 2018, 45-57
Available online at: http://journal.uny.ac.id/index.php/reid

Exploring the accuracy of school-based English test items for grade XI students of senior high schools

[*1]Martin Iryayo; [2]Agus Widyantoro
[1]University of Rwanda - College of Education
KG 11 Ave, Kigali, Rwanda
[2]Department of English Education, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
[*]Corresponding Author. E-mail: martiniryayo@gmail.com

Submitted: 08 June 2018 | Revised: 27 July 2018 | Accepted: 01 August 2018

Abstract

This study is set out to (1) explore the accuracy of school-based English test items developed by English teachers and (2) compare the relationship between the content covered by teacher and the students' success level. This research used the quantitative approach. The source of the data is all grade XI students' answers to the English test for the second semester of 2016/2017 academic year, and their English teachers' responses to the questionnaire. During this cross-sectional survey, 241 grade XI students and six English teachers were selected by using the total population sampling technique. To analyze the data, the IRT model was prioritized with BILOG MG 3.0, WINISTEPS 3.7. The findings of the study indicate that (1) the test is valid, (2) it is reliable, (3) majority of the items are moderately difficult, (4) more than a half of all items have power to discriminate the examinees, (5) some items show fully-effective distractors, and (6) the test gives much information at -.40 of theta which means that the test is difficult for the grade XI students. Moreover, there is a wide gap between the content covered and the level of success.

Keywords: *CTT, discrimination power, distractor, information function, IRT, theta, total population sampling*

Introduction

A well-constructed test is the best way to evaluate a student's mastery in a particular field. Gronlund (1993, pp. 205–206) stresses that tests do not only help teachers to make some instructional decisions with their direct influence on students' learning, but they also assist in a number of other ways. For instance, tests can increase students' motivation. The purpose of tests is to obtain an accurate and fair assessment of students' abilities. Nevertheless, it is impossible for a test to evaluate skills or knowledge bases if it is influenced by irrelevant factors that could undermine the results. These factors that potentially create bias can comprise of gender, ethnic, and cultural differences. In case there is no proper accounting for these biasing factors, the outcome of the test will unfairly represent the abilities of the examinees (Gronlund, 1993, p. 207). Alternatively, the results of a test are essentially meaningless if they are unfair for test takers due to the culture, gender, or ethnic origin biases.

A veracious picture of skills and knowledge that students have in either the subject area or domain tested should be presented by test results. The successful instructional, curriculum planning, and evaluation of linked programs cannot be accomplished without students' quality achievement data. Test scores that overestimate or underestimate students' actual knowledge and skills cannot serve these important purposes (Young, Cummings, & St-Onge, 2017). The accuracy of the achievement data cannot be procured since the composers of the test do not pay attention to the accuracy of the test components, because once the test is not well prepared, it obviously

ISSN 2460-6995

affects the students' achievements even if they understand the material well. Thus, the test must be as accurate as possible.

In standardized testing, there are several means for measuring students' cognitive abilities. Currently, multiple-choice tests are commonly used for measuring students' cognitive abilities (Galsworthy et al., 2005). Standardized scores is used by most schools to evaluate the educational quality and student performance (Brescia & Fortune, 1989, pp. 1–5). As long as it is believed that test scores are considered as an important factor to assess students' performance, teachers should develop the tests which are as fair as possible for examinees regardless of their races, genders, or any disability they may have (Joint Committee on Testing Practices of American Psychological Association, 2004). Reviewing all items of a test is the most fruitful way to ensure that they are free from all irrelevant sources of variances because item bias dirtily affects the examinees' scores. There must be empirical revision of the items before administering them to the examinees in order to ensure the quality of their characteristics.

In achievement testing, it is possible to use different formats. Multiple-choice (MC) items are broadly used for classroom assessment and they always account for a significant constituent of a student's grade in a course (DiBattista & Kurzawa, 2011). A normal MC item is made up of a question, known as stem, and a list of alternatives from which one becomes the right answer to the question. The test takers pick only the option they think fits to the question asked. The keyed option is the best name for the correct answer while the remaining alternatives refer to as distractors. For instructors, there is a variety of advantages to use the MC test format; scoring MC items takes a short time particularly when the examinees indicate their responses on a well-scanned optical MC answer sheet (universally used form). For teachers of subjects with large enrolment, easy grading can make MC tests very specifically appealing to them. Obviously, multiple choices tests are more advantageous even though there are some flaws still pending while measuring the students' performance.

Content validity, clarity, and reliability are the most crucial traits of achievement tests. The content validity of a test is always seen by how accurately the test samples the range of knowledge, skills, and abilities expected from the testees during an examination period. The reliability of a test depends on its grading stability and its power to discriminate students upon the basis of their different levels of performance (Kartowagiran, 2012). Well-developed multiple-choice test items are in general more valid, clearer, and more reliable than essay tests because they broadly represent content in the syllabus, able to distinguish all levels of performance, and scoring consistency is virtually guaranteed. Thus, validation is a starting point for dealing with multiple choice test item quality.

Content validity can be obtained in various ways. The content validity (relevance) by experts' judgement can be computed in different ways. The use of pre-established acceptability criterion, calculation of rating average upon each item relevance, quantification of item relevance (with three or more experts) by using coefficient alpha, and kappa coefficient computation are the most known techniques (Polit & Beck, 2006). With this approach, there should be a team of experts to judge whether an item on a scale is relevant to (or congruent with) the construct being measured. Each rater is free to compute the percentage of item relevance, then the average is taken across all raters (experts).

Another way of evaluating the accuracy of MC test items is concerned with studying the answers that the examinees make, in which within this research, this analysis approach was used. Precisely, teacher-developed test items administered to the examinees are basically analyzed on the basis of difficulty level, discrimination level, and effectiveness of the distractors (DiBattista & Kurzawa, 2011). In brief, before putting the items in their bank, the main characteristics stated above should be considered because any item which is either too difficult or easy, item that does not discriminate students, and item with ineffective distractors, does not qualify to be stored in the item bank.

Test takers should be differentiated by their abilities. The discrimination capacity of an MC test item is the most prevailing property because it reflects the extent to what more intelligent students are more likely than less knowledgeable students to select the keyed option (Abadyo & Bastari, 2015). MC test item discriminatory capacity can be measured with the computation of its index, which reflects the correlation between the examinees' total scores and the score received on the item to be considered (i.e. 1 stands for the keyed option, while 0 for the wrong answer). Even more, there are items which are problematic because they produce negative discriminatory indexes, maybe due to the unclear wording or the existence of two correct alternatives rather than one (DiBattista & Kurzawa, 2011). With the presence of such items, there is a detraction from the overall accuracy of the test as a whole, because the number of less knowledgeable examinees who select the keyed option outweighs that of the knowledgeable examinees.

With regard to the perspective of its functionality, there are two requirements for a distractor to be functioning: first, at least some examinees must select it, if they do not, the distractor is not plausible to them until they can be lured away from the correct answer, so such a distractor never contributes to the discrimination of the test takers. Abdulghani, Ahmad, Ponnamperuma, Khalil, and Aldrees (2014) have suggested that at least 5% of examinees should select each of an item's distractors, and this value is a common benchmark for the effectiveness of the distractors. The second requirement refers to the power of a distractor to distinguish high achievers from low achievers (stronger from weaker students), considering that the power of discrimination is clear when the correct answer is more often chosen by the students with high scores than their counterparts.

Related to the statements, opinions, and views of different authors as fully explained in the previous section, the problems that always appear when developing school-based English test items with the format of multiple choices, are so many, such as the content of some multiple-choice tests, which does not cover the material taught in the classroom, and the main parts of multiple choice tests items; stem, key, and alternatives which are not built according to the criteria or guidelines; some teachers do not have enough skills to get by this problem; by analyzing the scores of the students obtained from multiple choice tests during a couple of academic years ago, there is inconsistency because there is a lack of item homogeneity; some individual items are not highly correlated to each other and even to the whole test; some English teachers are not cautious of the difficulty level of the items. At the end of a teaching session, they develop tests which are either too easy or difficult. The ideal index of difficulty should fall between -2 and 2 (Hambleton, Swaminathan, & Rogers, 1991, p. 13). It is quite problematic to have items with difficulty index of far less than -2 or more than 2, and some English teachers do not know how to develop multiple choice items which can discriminate the participants. Moreover, the distractors are powerless to attract the examinees because some are chosen by <5% of examinees (Mkrtchyan, 2011). It is a problem to have items which cannot discriminate the achievers (a_i less than 2).

Like other scientific studies, this study aims at exploring the accuracy of school-based English test items developed by English teachers through (1) validity index, (2) reliability coefficients, (3) difficulty level, (4) discrimination power, (5) distractor effectiveness, and (6) level of information given by the items and the whole test in general and at comparing between the content covered by teacher and student success level. The current study is expected to be beneficial. Practically and even theoretically, the results of this study should be used by English test administrators, moderators, and even supervisors in order to make adequate policies on how to fairly and professionally prepare a suitable English test. This is very important because some teachers and other school academicians who develop test items for testing students do not have enough skills yet to examine the primordial characteristics indicating a good item.

Many researchers worked on the accuracy or quality of achievement test items.

Charismana and Aman (2016) conducted a research about the quality of civic education final examination items, in the whole regency of Kudus, Indonesia. The students involved in the study were grade VIII students of junior high schools that apply Curriculum 2013. The data were analyzed both qualitatively and quantitatively. The qualitative results show that 31 items are good whereas are items are not. The quantitative results show that 24 items or 68.57% of all items are good, while 11 items or 31.42% of all items are not. As a result, approximately 15 items are recommended to be revised.

A study conducted by Osadebe (2015) with 100 items administered to 1000 students comes up with the results that the achievement test for the subject of Economics has a high face and content validities. The test item quality was evaluated through difficulty and discrimination indexes. A difficult index or *p-value* of 0.5 was referred to after the use of the formula for guessing correction. The index of discrimination was computed with point biserial statistics whereby the minimum boundary is .30. With the KR-20, the test was very highly reliable with the coefficient of .95. These findings support the use of this instrument to internally evaluate the students in order to be ready for the external testing (examination).

According to the study by Boopathiraj and Chellamani (2013), which was aimed at analyzing test items in the subject of Research with students enrolled in Master of Education (M.Ed) program, they wanted to ensure the difficulty and discrimination levels of MC test items. A sample of 200 students from different colleges of education was established. The sample consisted of both genders. The findings indicate that a big number of items are not accepted, and there is a good discrimination index for some items, but some of them are rejected due to poor discrimination indexes. Based on the statement above, most of the items have the difficulty level (*bi*) from -2 to 2 and discrimination index of (a_i) > 2.

Sabri (2013) worked on a comprehensive test at a university in Perak, involving 16 music students. With MS Excel, he computed the difficulty level of 41 items. The reliability coefficients and discriminatory indexes were computed using MS Excel and SPSS 17.7 respectively. The outcome of the research came up with the information that 44% of all items have the difficulty index of > .80, then 59% of the items have acceptable discriminatory power. There is no effective distractor. With KR-20, the coefficient of reliability is .717 while with KR-21 is .703. Hence, it is reasonable to conclude that the items are reliable, moderately easy, 80% discriminate high from low achievers, but some distractors were chosen by less than 5% of examinees (implausible).

Quaigrain and Arhin (2017) carried out a study about MC test items. The sample was made up of 247 students doing year-1 diploma in education at Cape Coast Polytechnics. A test of 50 MC items was given to them in the subject of educational measurement. The results of the study show that the whole test has an internal consistency reliability of .77 (KR-20), the mean score of 29.23, the standard deviation mean score of 6.36, difficulty level (*p-value*) and discrimination index (*DI*) of 58.46% (SD=21.23) and .22 (SD=.17), respectively, and the mean score of DE of 55.04 (SD=24.09). As to DI, 30 items (60%) are reasonably accepted. Every item with moderate difficulty level, high discriminatory power, and functioning distractors should still be part of the next testing to improve classroom assessment quality.

There is no study without innovation. The novelty of this study can be seen from data analysis section. Apart from the variables that look similar to the previous studies by other academic researchers, the current research involves a new way of giving grades to teachers on the basis of content covered after the learning term. As the majority of the previous studies used classical test theory to analyze item accuracy, the researchers in this study used the item response theory (IRT) to have clearer and more information on the item quality, so that the newly published IRT software was used.

This study is expected to come up with the answers to the questions in relation to the quality or accuracy of school-based English test items: (1) To what extent do English test

items represent the content or subject topics they intend to measure for grade-11?; (2) What proves that English test assesses the underlying theoretical construct it is purported to measure?; (3) How convergent are the items, making up English test, to be considered homogeneous? Do they complement each other?; (4) How reliable and informative are the English test items?; (5) What is the difficulty level of the items making up English test?; (6) At what level do English test items powerfully discriminate between high and low achievers?; How effective are the distractors to ensure that English test outcomes provide more credible and objective picture of the knowledge of the examinees?

Method

This study used the quantitative approach with a cross-sectional survey. It was carried out within the period of two months, from the end of May to the mid-June 2017. The study took place across all senior high schools under the management of Muhammadiyah foundation. The schools are situated in Bantul District, Special Region of Yogyakarta, Indonesia. In order to successfully reach the objectives of this study, the schools which are homogeneous were considered.

Population and Sample

The population of this study was all Muhammadiyah high school students of grade-11 in the whole district of Bantul, totalling 241 students. In order to have accurate results, all of the students were selected as participants. By the small community, it is possible to conduct a study with nearly the whole population and pay attention to whoever has moved through the network of the community (Guyette, 1983). Therefore, this study uses the purposive sampling technique with total population sampling.

Data Collection Techniques

The technique used for data collection is documentation whereby the researchers recorded the answers from all examinees. To have information on the content covered by each teacher during the learning session, a questionnaire was used. With regard to the validity and reliability of the instruments in this study, experts' judgement and Crobach's Alpha indexes were computed.

Data Analysis

Within the scope of this study, there are a lot of variables to be measured, including construct validity, internal consistency reliability, item level of difficulty, the level of discrimination, and the effectiveness of the distractors. It is, therefore, clear that both Classical Test Theory (CTT) and Item Response Theory (IRT) are necessary in this analysis. Table 1 displays the variables and related data analysis techniques.

Table 1. Data analysis techniques

No	Variables	Analysis Techniques
1	Validity: Content Validity	Expert judgments with Aiken indices
2	Reliability: Internal Consistency Information Function	 JASP 0.8.2.0 = SPSS24 IRT/BILOG-MG 3
3	Level of Difficulty	IRT/ BILOG-MG 3
4	Power of Discrimination	IRT/ BILOG-MG 3
5	Distractor Effectiveness	Rasch/WINISTEPS 3.73

Table 1 contains the variables of the study and the analysis related to them. The coefficient of reliability which can be accepted must have a minimum of .70. This value helps to determine the level of error within measurement. The higher the index of reliability is, the higher the level of errors within measurement decreases, and vice versa (Mardapi, 2012, p. 128). Item discrimination (a_i) is the power of an item, by which its score is used for differentiating the examinees whose level of understanding is high from those whose level of understanding is low. The discrimination index is called *slope* because it shows the extent to which the probability to change the correct response like the ability or increase of the trait exists. According to Hambleton and Swaminathan (1985, p. 36), discrimination index varies from 0 to 2.

The item difficulty is another important variable. Its index (b_i) is always measured from the scores of students or examinees which are

obtained from the answers of all participants in a test. Item difficulty depends on the ability of the examinees. The more the testees have correct answers on an item, the higher the difficulty level of that item flops or decreases and vice versa. The item which is good or accepted is always situated between the interval of $-2 \leq \theta \leq 2$ (Hambleton et al., 1991, p. 13). The level of difficulty decreases as the *b*-parameter value is close to -2, but when the *b*-parameter value is close to +2, the level of difficulty increases.

The item analysis by using IRT model must fulfills the prescribed assumptions. The general assumptions that always appear in Item Response Theory models are unidimensional, local independent, and invariant parameters. The proof of unidimensionality is proven by the plot called Scree Plot as presented in Figure 1.

Figure 1 shows that a unidimensional assumption is fulfilled for this study data analysis because there is one most dominant dimension. The only way to test the model fitness is statistical measurement with chi-square. The researchers chose the suitable model by considering the highest percentage as shown by Figure 2 (Stone, Ye, Zhu, & Lane, 2009).

Figure 2 shows that the data in this study fit more to the second parameter model because it contains 36% (18 of 50) of all items. This result also supports the invariance assumption because when the data fit a model, the invariance criteria are automatically fulfilled (Lord, 2012, p. 126).

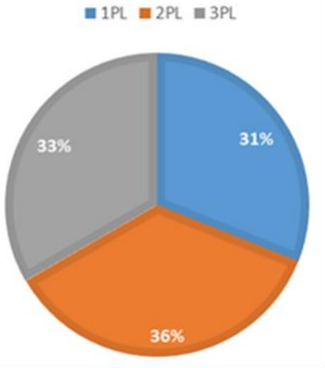

Figure 2. Goodness of fit (GoF)

The local independence has two facets: the local independence towards the test takers' answers and local independence towards the test items (Allen & Yen, 2001, p. 241). The first facet means that the wrong or right answer of a test taker does not depend on the wrong or right answer of his/her co-test taker on a given item. The second facet means that to be wrong or right on a test item does not affect the answer to another item. This study puts interest on the second facet of local independence because it is related to the test items. The results show that the correlation of residuals for all items is close to 0.

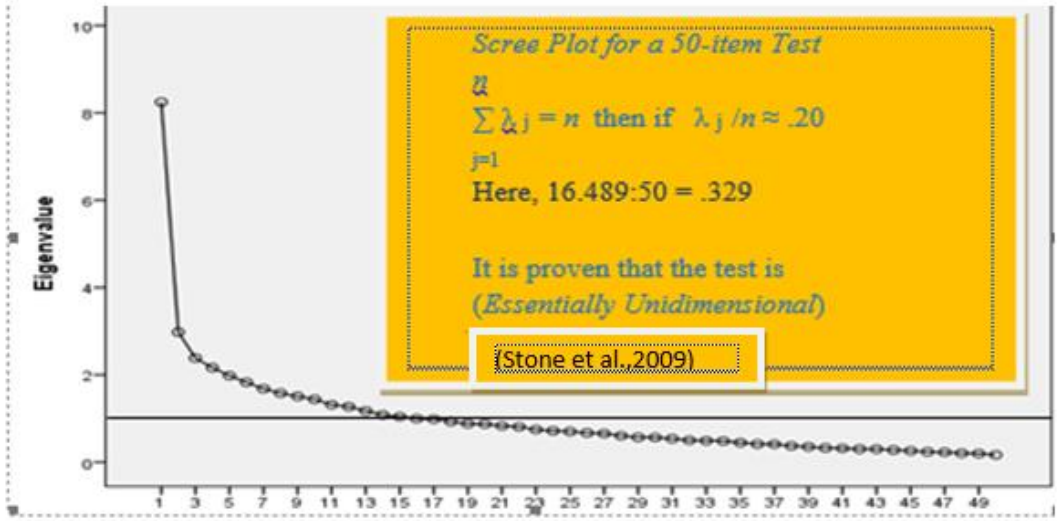

Figure 1. Unidimensionality proof by scree plot

Findings and Discussion

The findings of this study are discussed based on the variables to be measured. Validity index, reliability coefficient, discrimination index, difficulty level, distractor effectiveness, information function, and the success and content coverer weight were measured and the results can be found in this section. The findings about content validity with Aiken index are displayed in Figure 3.

Figure 3 contains information about the validity of items developed from the content expected to be covered by English teachers. It is supported that the English test represented the content taught because the Aiken Index for each indicator is accepted with the value bigger than .75. All items should be used because the overall index is .80. This result is supported by (Retnawati, 2016) who states that if the index is lower than or equal to .40, the validity is still low, if it is between .40 and .80, the validity is moderate, and if it is >.80, the validity is very high.

Reliability is another important criterion for item accuracy. Table 2 shows how reliable each item is. The Guttmann's Lambda$_7$ is the alternative of Cronbach's Alpha. Both coefficients were used to make a comparison.

The reliability coefficients are really good. Based on both Cronbach's and Guttmann's indices, the values range from .80 to .95. All items are perfectly reliable because any item's reliability greater than .70 is considered perfect, and the lowest and highest boundaries are .00 and 1.0 respectively. With this finding, there is no doubt that the students' answers to each item of the test are consistent. Hence, the test was measuring what it was purported to measure.

Table 2. Internal consistency reliability

Item	α	λ6	Item	α	λ7
Item1	0.88	0.92	Item26	0.89	0.93
Item2	0.88	0.92	Item27	0.88	0.92
Item3	0.88	0.92	Item28	0.88	0.92
Item4	0.88	0.92	Item29	0.88	0.92
Item5	0.88	0.92	Item30	0.88	0.92
Item6	0.88	0.93	Item31	0.88	0.92
Item7	0.88	0.92	Item32	0.88	0.92
Item8	0.88	0.92	Item33	0.88	0.92
Item9	0.88	0.92	Item34	0.88	0.92
Item10	0.88	0.92	Item35	0.88	0.92
Item11	0.88	0.92	Item36	0.88	0.92
Item12	0.88	0.92	Item37	0.88	0.92
Item13	0.88	0.92	Item38	0.88	0.92
Item14	0.88	0.92	Item39	0.88	0.92
Item15	0.88	0.92	Item40	0.88	0.92
Item16	0.88	0.92	Item41	0.88	0.92
Item17	0.88	0.92	Item42	0.88	0.92
Item18	0.88	0.92	Item43	0.88	0.93
Item19	0.88	0.92	Item44	0.88	0.92
Item20	0.88	0.92	Item45	0.88	0.92
Item21	0.88	0.92	Item46	0.88	0.92
Item22	0.88	0.92	Item47	0.88	0.92
Item23	0.87	0.92	Item48	0.88	0.92
Item24	0.88	0.92	Item49	0.88	0.92
Item25	0.88	0.92	Item50	0.88	0.92

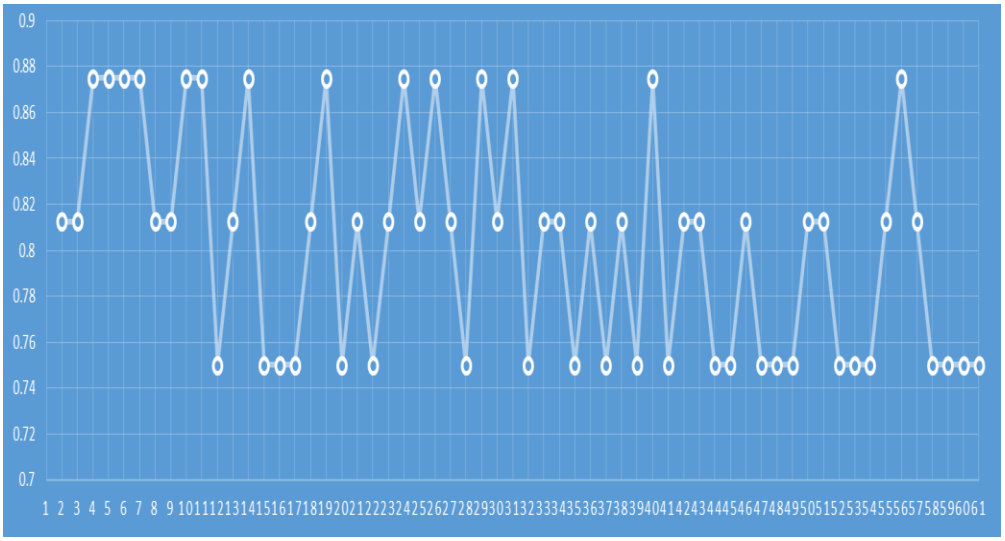

Figure 3. Aiken index (0.0 to 1.0)

The level of difficulty is very crucial to ensure the quality of test items. The results of *b*-parameter estimation for all English test items are summarized in Table 3.

Table 3. Difficulty index (*bi*)

Comment	Frequency	%
Good	47	94
Not good	3	6
Total	50	100

The parameter estimation for all 50 items shows that only three items (6%) are classified 'not good'. Those items are items 1, 40, and 46. The classification of item difficulty index relies on the range varying from -2 to 2 (good), and if it is out of the range, then it is not good. This result is in line with Mardapi (1991, p. 11) who states that the item difficulty level is the function of the ability of a test taker. An item is said to be good if it has the difficulty level (*bi)* between $-2 \leq b \leq +2$. An item with the difficulty level close or below -2 shows that the item is in an easy category. In contrary, an item with difficulty level (*bi*) close or above +2 shows an item that is in a difficult category. Figure 4 shows more about the accuracy of the test items based on *b*-parameter. The diagram in Figure 4 shows the test level of difficulty:

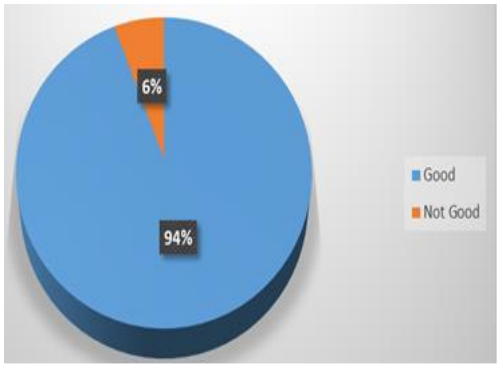

Figure 4. Item accuracy based on bi-index

Apart from the difficulty level, test items must be able to discriminate students by their abilities. The discrimination index for each item out of 50 items is well indicated in Figure 5. In terms of the discrimination index (*ai*), items 5, 10, 24, 35, 43, and 47, (12%) discriminate test takers at a low level because their *a-indexes* vary from between *.35 to -.64*. Items 6, 16, and 27, (6%) discriminate the examinees at a very low level because their *a-indexes* vary from .01 to .34.

However, the overall *a-index, 1.206,* shows that the English test moderately discriminates the examinees. Hence, all items with low discrimination indexes should be revised, while those with very low discrimination index should be replaced. The results are well shown in the diagram in Figure 5.

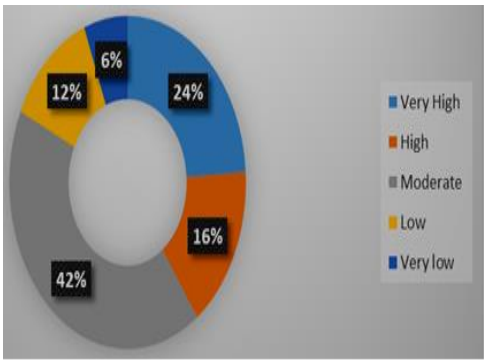

Figure 5. Discrimination power

The results presented in Figure 5 are supported by Baker (2001, p.34) that Discrimination Index (*ai*):

0.01 – 0.34 very low;
0.35 – 0.64 low;
0.65 – 1.34 moderate;
1.35 – 1.69 high;
1.70, and above very high.

Discrimination index (*ai*) is connected to the distractors' power to attract the examinees. The results can be seen in the diagram in Figure 6.

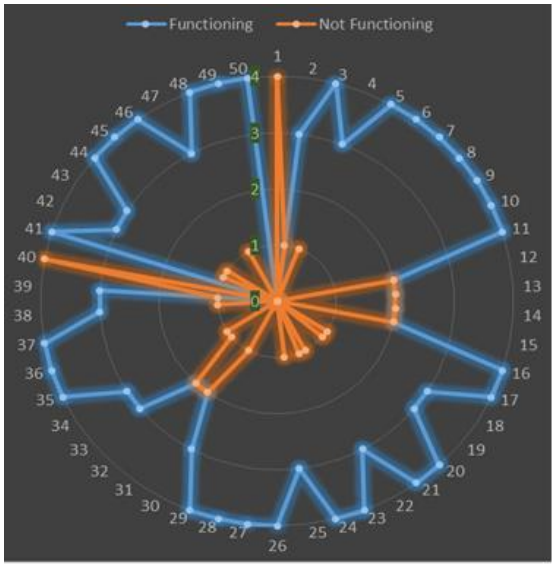

Figure 6. Distractor functionality

Notes:
0-4: Number of distractors (functioning ≥ 5% or not functioning ≤ 5%)
1-50: Number of items

It was found that items 1 and 40 (4%) do not have any functioning distractor, items 12, 13, 14, 15, 31, and 32 (6 items, 12%) have 50% of distractors that are not functioning effectively, items 2, 4, 18, 19, 22, 23, 25, 30, 33, 34, 38, 39, 42, 43, and 47 (15 item, 30%) have 25% of distractors that are not functioning, and items 3, 5, 6, 7, 8, 9, 10, 11, 16, 17, 20, 21, 24, 26, 27, 28, 29, 35, 36, 37, 41, 44, 45, 46, 48, 49, and 50 (27 items, 54%) have distractors that are functioning at 100%. In general, the English test for grade XI students during the second semester of the academic year of 2016/2017 has only 27 perfect items, two items that should be removed, and 21 items that should be repaired. Figure 6 represents the power of distractors within the test. These findings are supported by Abdulghani, Ahmad, Ponnamperuma, Khalil, and Aldrees (2014) who suggest that at least 5% of examinees should select each of an item's distractors, and this value is a common benchmark for the effectiveness of distractors.

The information function is another indicator of test item accuracy. In the IRT, the information function stands for the reliability. In this study, the plot was used to easily see the amount of information the test could give, as presented in Figure 7.

The maximum information can be seen on the student's ability of -.04. On the other hand, the red line shows the error of measurement (SEM), the more information line picks, the fewer the error of measurement values drops. In fact, the majority of grade XI students have a low ability because the test gives much information on the left side from 0 on the latent trait. We can see that the test is fit for the students whose abilities vary from -.2.2 to 1.4. This is supported by Istiyono, Mardapi, and Suparno (2014).

As seen in Figure 8, around 70% (169 students) of all students (241) have a low ability to answer the questions. Therefore, there is no easy item for the students because their abilities are relatively low, -.40.

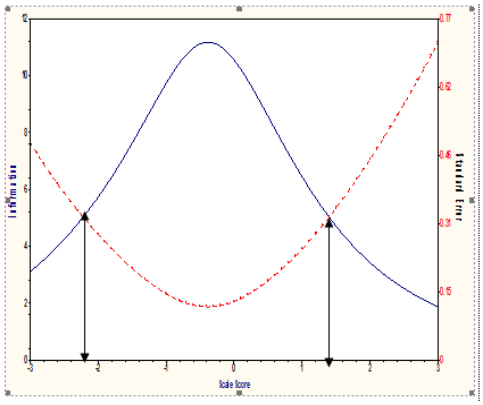

Figure 7. Information function (IF)

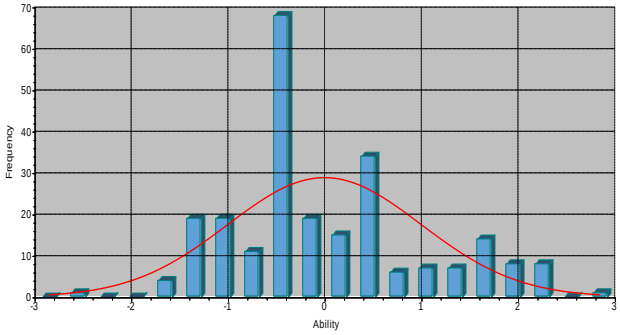

Figure 8. Proportition of students' abilities

As previously described, the content covered and the level of success were compared to see whether the students could understand that content well. Table 4 and Figure 9 have much information on the issue.

From Table 4, it is easy to evaluate, compare, and classify teachers at the end of a teaching term/period. It is known that every teacher has a syllabus that encloses the whole material. Every English teacher has the objectives to be achieved by the end of the term. A is given to an English teacher who reaches the target, B for a teacher who reaches an acceptable level, C for a teacher who needs to improve his/her teaching topics, D for a teacher who does not cover the content to the satisfaction, and F to a teacher that does cover a very minimum content. A= 82.5 to 100% of the content covered, B= 62.5 to 82.4% of the content covered, C= 42.5 to 62.4% of the content covered, D= 22.5 to 42.4% of the content covered, and F= 20% and below the content covered.

Apart from the teacher categorization criteria above, the new teacher project, as cited by Seidel, Stürmer, Blomberg, Kobarg, and Schwindt (2011), suggests a way to give scores to teachers. In the report called *Rating a Teacher Observational Tool,* the teachers can be put into categories, including: 'complete coverage' when the tool of evaluation covers all the elements in the curriculum, 'partial coverage' when the test does not cover some components of the syllabus, and 'inadequate coverage' when the evaluation tool covers lower than 50% of all indicators in the syllabus. Figure '3' stands for the first category, '2' for the second, and '1' for the third. Based on the answers of the teachers, all six teachers were categorized.

With Figure 9, it is easy to see the gap between the content covered and the success level of grade XI students. There are some English teachers, ENGT.BL, ENGT.SW, ENGT.PL, who show that content and success are in line, but the rest of the teachers, ENGT.PY, ENGT.IM, ENGT.KS, indicate a long gap between the content covered and success of students on the English test. Information from Figure 9 implies that there is a remarkable difference between rural and urban Muhammadiyah senior high schools. For the rural schools, the content covered by English teachers does not explain the success level of students on the test developed from that content, but for the urban schools, there is correlation between the content covered and the success level of the students.

Table 4. Classification of English teachers

ID CODE	Indicators Covered/61	Scale	Grade	Comment	Category	Comment
ENGT.BL	53.00	3.5	A	Reached Target	2	Partially Covered
ENGT.PY	53.00	3.5	A	Reached Target	2	Partially Covered
ENGT.IM	52.00	3.4	A	Reached Target	2	Partially Covered
ENGT.SW	49.00	3.2	B	Acceptable	2	Partially Covered
ENGT.KS	44.00	2.9	B	Acceptable	2	Partially Covered
ENGT.PL	37.00	2.4	C	Need Improvement	2	Partially Covered

Figure 9. Content covered vs success

Conclusion, Implications, and Suggestions

In connection with the results of this study and its discussion within the previous chapter on the accuracy of the multiple-choice items of the English test, different concluding statements can be made as follows: (1) The items represent the content taught to the students during the second semester of academic year 2016/2017. (2) All items are internally consistent. (3) Most of the items (47/50) have acceptable difficulty level, but there are two items which are very easy and one which is very difficult. (4) A big number of items (42/50) have good discrimination indexes, but nine items are unable to discriminate the high achievers from low achievers. (5) As many as 27 items have effective distractors, but 23 items still show powerless distractors. (6) Some English teachers tried their best to cover the content expected to be taught to the students, but some others did not cover at least 50% of the content, and therefore, there is still a gap (for some schools) between the content covered and the success level of the students on the test developed from that content. (7) The test is obviously difficult for more than 70% of the students who have the ability of -.40, and fits the students whose abilities range from -2.2 to 1.2.

Like in other scientific studies, some implications are put forward that the improvement in constructing and developing English test items for grade XI students of Muhammadiyah senior high schools in Bantul district needs both qualitative and quantitative review. It is necessary to test the quality of each item. This process contributes to the identification of some weaknesses within the test because the quality level of a test is completely determined by the quality of its items.

The results of the quantitative analysis of the English test, in general, are not accurate. The teachers should make some try outs of the items, then the results are analyzed with relevant and practical techniques, such as the item analysis with the classical test theory and item response theory as well. The determination of the technique of analysis depends on the purpose and number of examinees accompanied by other technical assessments. An analysis with the classical test theory needs a small sample (30 participants at minimum), but the item response is used for a big number of respondents.

For a better future school-based assessment, the following suggestions are given: (1) All items with medium quality should be revised, re-measured until they fulfill the criteria of a good item; the items with bad quality should be dropped or completely replaced. (2) It is much better for the teachers to conduct some tryouts and analysis of items before testing. (3) It is quite advisable for the teachers to develop items that are suitable to the content that is already taught to the students; they should also give the blueprint to them. (4) Before a set of items are chosen, it is necessary to conduct qualitative analysis with expert judgment. It can help English teachers to have information on the item characteristics in terms of construction, language, and content in general. (5) The item response theory is needed to identify the characteristics of items; IRT related programs should be trained to teachers of senior high schools. (6) It is suggested to make a test item bank at the district level (Bantul) for the English subject to help teachers practice in assessing students' achievements. (7) Schools should prepare some routine trainings on evaluation, assessment, and measurement. It will help to increase the ability English teachers in evaluating learning outcomes. The management office of Muhammadiyah schools should be vigilant to remote areas in terms of education and technology.

References

Abadyo, A., & Bastari, B. (2015). Estimation of ability and item parameters in mathematics testing by using the combination of 3PLM/ GRM and MCM/ GPCM scoring model. *REiD (Research and Evaluation in Education)*, *1*(1), 55–72. https://doi.org/10.21831/reid.v1i1.4898

Abdulghani, H. M., Ahmad, F., Ponnamperuma, G. G., Khalil, M. S., & Aldrees, A. (2014). The relationship between non-functioning distractors and

item difficulty of multiple choice questions: A descriptive analysis. *Journal of Health Specialties*, *2*(4), 148–151. https://doi.org/10.4103/1658-600X.142784

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory* (1st ed.). Long Grove, IL: Waveland Press.

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* (Vol. 2).

Brescia, W., & Fortune, J. C. (1989). Standardized testing of American Indian students. *College Student Journal*, *23*(2), 98–104.

Charismana, D. S., & Aman, A. (2016). Analisis kualitas tes ujian akhir semester PPKN SMP di Kabupaten Kudus. *Jurnal Evaluasi Pendidikan*, *4*(1), 1–9.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), 1–23. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Galsworthy, M. J., Paya-Cano, J. L., Liu, L., Monleón, S., Gregoryan, G., Fernandes, C., … Plomin, R. (2005). Assessing reliability, heritability and general cognitive ability in a battery of cognitive tasks for laboratory mice. *Behavior Genetics*, *35*(5), 675–692. https://doi.org/10.1007/s10519-005-3423-9

Gronlund, N. E. (1993). *How to make achievement tests and measurements*. Needham Heights, MA: Allyn and Bacon.

Guyette, S. (1983). *Community-based research: A handbook for native Americans*. Los Angeles, CA: American Indian Studies Center, University of California.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (PhysTHOTS) peserta didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. https://doi.org/10.21831/pep.v18i1.2120

Joint Committee on Testing Practices of American Psychological Association. (2004). *Code of fair testing practices in education*. Washington, DC, United States of America.

Kartowagiran, B. (2012). *Penulisan butir soal*. A paper presented in the Seminar on Question Items Analysis and Writing for Civil Servant Resources of Dik-Rekinpeg, in Kawanua Aerotel Hotel.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.

Mardapi, D. (1991). Konsep dasar teori respons butir: Perkembangan dalam bidang pengukuran pendidikan. *Cakrawala Pendidikan*, *3*(X), 1–16.

Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.

Mkrtchyan, A. (2011). Distractor Quality Analyze In Multiple Choice Questions Based On Information Retrieval Model. *EDULEARN11 Proceedings*, 1624–1631.

Osadebe, P. U. (2015). Construction of valid and reliable test for assessment of students. *Journal of Education and Practice*, *6*(1), 51–56.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent*

Education, *4*(1), 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian*. Yogyakarta: Parama Publishing.

Sabri, S. (2013). Item analysis of student comprehensive test for research in teaching beginner string ensemble using model based teaching among music students in public universities. *International Journal of Education and Research*, *1*(12), 1–14.

Seidel, T., Stürmer, K., Blomberg, G., Kobarg, M., & Schwindt, K. (2011). Teacher learning from analysis of videotaped classroom situations: Does it make a difference whether teachers observe their own teaching or that of others? *Teaching and Teacher Education: An International Journal of Research and Studies*, *27*(2), 259–267. https://doi.org/10.1016/j.tate.2010.08.009

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2009). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*(1), 63–86. https://doi.org/10.1080/08957340903423651

Young, M., Cummings, B.-A., & St-Onge, C. (2017). Ensuring the quality of multiple-choice exams administered to small cohorts: A cautionary tale. *Perspectives on Medical Education*, *6*(1), 21–28. https://doi.org/10.1007/s40037-016-0322-0