Indonesian Journal of EFL and Linguistics
Vol. 5 No. 2, 2020
eISSN: 2503-4197, pISSN: 2527-5070
www. indonesian-efl-journal.org

Item Analysis of English Final Semester Test

Amalia Vidya Maharani
Yogyakarta State Univeristy
e-mail: ammelvm@gmail.com

Nur Hidayanto Pancoro Setyo Putro
Yogyakarta State University
e-mail: nur_hidayanto@uny.ac.id

Abstract:
Numerous studies have been conducted on the item test analysis in English test. However, investigation on the characteristics of a good test of English final semester test is still rare in several districts in East Java. This research sought to examine the quality of the English final semester test in the academic year of 2018/2019 in Ponorogo. A total of 151 samples in the form of students' answers to the test were analysed based on item difficulty, item discrimination, and distractors' effectiveness using Quest program. This descriptive quantitative research revealed that the test does not have good proportion among easy, medium, and difficult item. In the item discrimination, the test had 39 excellent items (97.5%) which meant that the test could discriminate among high and low achievers. Besides, the distractors could distract students since there were 32 items (80%) that had effective distractors. The findings of this research provided insights that item analysis became important process in constructing test. It related to find the quality of the test that directly affects the accuracy of students' score.
Keywords*: distractor, item analysis, item difficulty, item discrimination*

1. INTRODUCTION

A large number of studies have highlighted the crucial roles of appropriate assessment in the success of English teaching-learning process. The success of English teaching-learning can effect on students' language proficiency. Sulistyo and Suharyadi (2018) argue that students who have well language proficiency can use that language to communicate. Achieving that success, assessment can provide student progress in mastering the material that has been given (Browder et al., 2006). It assists the teacher to determine the proper approach and method of teaching (Scouller, 1998). As stated in *Peraturan Menteri Pendidikan Dan Kebudayaan RI Tentang Standar Penilaian Pendidikan* (2016), assessment as a process of collecting and processing information to measure the achievement of students' learning outcomes in learning activity. Instrument is needed in obtaining information of assessment. It is a process where information is produced to oversee the improvement of students' abilities. (Arikunto, 2016) identified two types under the term assessment which can use as an instrument: tests and non-tests. Tests involve diagnostic, formative, and summative test. Whereas non-tests involve rating scale, questionnaire, checklist, interview, observation, and biography.

Generally, Indonesian teachers apply tests, specifically summative tests to assess students in the end of learning process. Brown (2004) states that test is a set of equipment to measure an individual's proficiency within particular criteria. This definition is close to that of (Miller et al., 2009) who define test as an equipment to assess students' abilities through a package of questions within a specified time. It means a test assists teachers to evaluate students' competence that can interpret students' progress. Furthermore, (Brown, 2004) explains the summative tests itself can assist teachers to assess students' comprehension when the learning process ends. It is one of the ways to discover the students' competencies in the end of learning process in the school. Automatically, teachers should construct a good test.

A good test needs to consist of well-constructed items which in turn will teachers to assess students' competencies accurately. It should consist of at least three criteria encompasses practicality, reliability, and validity (Brown, 2001). Practicality can broadly be defined as operating budget, time limitation, implementation, and scoring system of test. Test should be prepared with the low budget (Brown, 2001). Then, test should have vivid time limitation and could be managed easily. The most important is spelling out specific and efficient scoring system. Associating with reliability, the test result should provide stable results in different circumstances (Flucher & Davidson, 2007). Therefore, the test result is trusty. Whereas reliability refers to dependability, validity refers to the tests' ability to measure what should be measured accordance with the learning goals or competencies to be achieved. In ensuring the test has good quality, it must be analyzed to identify the quality by doing item analysis.

Previous studies have been conducted on the quality of English final semester tests, specifically in junior high school in Indonesia (for example Amelia, 2010; Maghfiroh, 2010; Toha, 2010; Ani, 2011; Lestari, 2011; Risydah, 2014; Haryudin, 2015; Manfenrius et al., 2015; Fajriah, 2016; Haryudin & Santosa, 2016; Pradanti et al., 2018;

Maghfiroh, 2019). However, none studies conduct item analysis of English final semester tests for junior high school in Ponorogo district. The interview results with the English teachers of junior high school in Ponorogo also showed that they often pass analyzing test items before distribute the tests to students.

Therefore, the purpose of this study is to describe the quality of English final semester tests for nine grade students in the academic year of 2018/2019 in Ponorogo district in terms of item difficulty, item discrimination, and effectiveness of distractors. These characteristics have been chosen partly because of the English teachers' forum of junior high school in Ponorogo district already analyzed theoretically. This study is expected to provide a feedback and an example for English teachers, educators, test developers, and others who create an English test. In addition, this study is done to provide a reference for future similar study.

2. LITERATURE REVIEW

2.1 Test Item Analysis

Assessment is a process where information is produced to oversee the improvement of students' abilities. For Miller et al. (2009), assessment means mechanism to find out the students learning results and progress through observation, projects, and tests. Researchers were pointed out in the previous that English teachers conduct summative test to assess the students' competencies in the end of learning process. Teachers or test makers should construct a good test so that the results are valid and reliable. In terms of a good test, Mardapi (2015) states nine steps for creating a highly qualified test involve: (1) composing test specifications, (2) creating a test, (3) analyzing a test, (4) doing a trial, (5) analyzing test items, (6) correcting test, (7) assembling test, (8) administering test, and (9) interpreting test results. Following those steps will assist teachers or test makers generating a well-constructed test.

As the interview results, the English teachers' forum of junior high school in Ponorogo district does not conduct analyzing test items before distribute the test. Test item analysis is claimed as the process to identify the quality of test. Rosana and Setyawarno (2017) say that item analysis is a method to dig up the test quality in order to refine the well-constructed item. In short, it is organized to identify and analyze the quality of test items. The major purpose of this process is to build on the better tests by revising or dropping poor items (Boopathiraj & Chellamani, 2013; Mukherjee & Lahiri, 2015). This process is important to confirm well-constructed items that are fit with the test principles. Moreover, teachers or test makers' ability in constructing test items will improve. The teachers or test makers role are revising or dropping test items that are not proper.

In analyzing test items, a good test at least should conform to three characteristics, namely item difficulty, item discrimination, and effectiveness of distractors (Brown, 2004). This is done by analyzing the students' responses of each item. Test makers can analyze by two statistical theories, namely classical test theory (CTT) and item response theory (IRT) (Haladyna, 2004). Item response theory is provided as a development of classical test theory. In classical test theory, the item difficulty index depends on the

number of samples. Otherwise, item response theory has advantage of providing estimation of difficulty appropriate to estimation students' ability (Flucher & Davidson, 2007). Since the researcher identify item difficulty, item discrimination, and effectiveness of distractors, this study used classical test theory. Classical test theory assumes that the assessment instrument has none errors which result in the participants have a true score.

Relating to this study, the researchers use classical test theory by Quest program. Quest program is one of computer-based statistics programs from The Australian Council for Educational Research Limited (ACER) (Izard, 2005). This program can increase the precision of calculation compared to the manual technique. Ofianto (2018) adds that Quest program can calculate by Classical Test Theory (CTT) and Item Response Theory (IRT). It means this program has advantages compared to other computer-based statistics programs. Suyata (2016) mentions the others benefits of Quest program are more accurate than other statistic programs. In addition, this program can analyze polytomous, dichotomous, and combination of dichotomous and polytomous data.

The TPAtn file output in Quest program displays about item difficulty, item discrimination, and effectiveness of distractors index. The item difficulty index is served as a value percentage that has an asterisk symbol. Further, the discrimination index is served from biserial point that has an asterisk symbol. Meanwhile, the distractor is served from the percentage of participants who choose the option. The options of being the distractor must have a lower biserial point than the correct option.

2.2 Item Difficulty

The item difficulty is to identify the percentage of students who answer correctly (Haladyna, 2004). This definition is similar to that found in Brown (2004) who writes: item difficulty relates to the percentage of students who assume an item easy or difficult. This characteristic identify whether the item is difficult or easy so this characteristic can assist the teachers in analyzing easy, medium, and difficult item. Kunandar (2013) claims that a test package must contain 25% easy items, 50% moderate items, and 25% difficult items. It will reduce students to become discouraged and not enthusiastic in answering test items. Arikunto (2016) argues difficult items cause students to be lazy in answering the questions.

The requirement that an item has an ideal item difficulty is that an item must neither too easy nor difficult. The range of item difficulty index is between 0.0 and 1.0. According to Flucher and Davidson (2007), the item difficulty index is between 0.30 and 0.70. Items with index less than 0.30 mean difficult while items with index more than 0.70 mean easy. Factors which affect item difficulty are item analysis theories, the clarity of questions, and similarity between test items with materials in syllabus (Haladyna, 2004).

Numerous studies have attempted to explain the item difficulty in relation to analyze the tests quality (for example Amelia, 2010; Maghfiroh, 2010; Ani, 2011; Risydah, 2014; Haryudin, 2015; Manfenrius et al., 2015; Maghfiroh, 2019; Pradanti et al., 2018). Some analysts, (e.g. Amelia, 2010; Ani, 2011; Maghfiroh, 2010; Risydah, 2014; Pradanti et

al., 2018; Maghfiroh, 2019) have attempted to analyze the item difficulty of English final semester test of junior high school. Thus far, these previous studies have revealed that the moderate items are more than others categories item. In summary, those test packages have more items that qualify as a well-constructed item than qualify as a poor-constructed item. Nevertheless, the portion among easy, moderate, and difficult items is not balanced.

In contrast to those six previous studies, Haryudin (2015) found that the difficult items are more than other categories item. In their analysis of item difficulty, these previous researches point out that those test packages have more poor-constructed items than well-constructed items. Different finding exist in the research regarding item difficulty analysis. Manfenrius et al. (2015) analyzed three test packages from three junior high schools. In their research, six items from 150 items were classified as difficult item. Most items were classified as easy item. Moreover, the portion between easy, moderate, and difficult items in this research is far from ideal.

An important theme emerges from the researches discussed so far: the ideal portion between easy, moderate, and difficult items. It is a challenge for teachers or test makers to create items with balanced portion. Thus, the items truly assist teachers to test their students based on students' ability.

2.2 Item Discrimination

Second characteristic is item discrimination that have ideal index more than 0.39 (Ebel & Frisbie, 1991) with range between 0.0 and 1.0 (Hingorjo & Jaleel, 2012). This characteristic is about identifying students' knowledge and ability (Haladyna, 2004). It assists teachers to discover high achievers and low achievers in a class. An item test can reach ideal index when high achievers answer correctly more often than low achievers (Hingorjo & Jaleel, 2012). However, this characteristic depends on the number of students' responses, which test makers analyze (Flucher & Davidson, 2007). This number of sample illustrates test takers' abilities. The smaller number of responses causes inaccurate of the item discrimination calculation. Another significant effect of item discrimination is the poor item discrimination index will give bad effect on reliable interpretation of the real students' knowledge (Setiyana, 2016).

Much of the previous studies emphasize the item discrimination analysis (for example, Toha, 2010; Lestari, 2011; Risydah, 2014; Haryudin, 2015; Manfenrius et al., 2015; Fajriah, 2016; Haryudin & Santosa, 2016; Pradanti et al., 2018; A. Maghfiroh, 2019). A number of authors have reported that less than 50% of the items are very good items to discriminate the high and low achievers (Toha, 2010; Risydah, 2014; Fajriah, 2016; Haryudin & Santosa, 2016; Pradanti et al., 2018; A. Maghfiroh, 2019). In contrast, different finding has been found to be related to item discrimination analysis (Lestari, 2011). Finding from this study presented that more than 50% of the items qualified as very good discrimination index. Conversely, other studies (see Haryudin, 2015; Manfenrius et al., 2015) reported that no items qualified as very good item discrimination.

In view of all that has been mentioned so far, one may suppose that the existing test items on those previous studies are not be able to discriminate high and low achievers. As Pradanti et al. (2018) argue that teachers or test makers must create items using vivid instructions and language structures. It can prevent students from confusion and difficulty while finishing the test.

2.3 Effectiveness of Distractors

Another characteristic is distractors. This characteristic can only be analyzed on tests in the form of multiple-choice tests. A well distractor must be chosen by at least 5% of the respondent, especially those who include in low achievers (Rosana & Setyawarno, 2017). In doing item analysis, test makers must analyze the effectiveness of distractors to measure the functioning incorrect options in attract students (Brown, 2004). Distractors analysis is one of important parts since it has several functions in item analysis. The functions involve reducing items that use ineffective sentences or too many options, providing information to improve the items, assisting to choose a correct distractor, assisting to comprehend students' cognitive behavior, and increasing items' response score (Haladyna, 2004).

Previous researchers have identified the effectiveness of distractors in tests. Several researchers have reported that the effectiveness of distractors in their studies is low (e.g. Risydah, 2014; Haryudin, 2015; Manfenrius et al., 2015; Pradanti et al., 2018; A. Maghfiroh, 2019). Data from these studies identified that more than 40% items qualified as ineffective distractors. Considering all of this evidence, it seems that teachers or test makers should increase their ability in constructing test items. The unclear language structures and unfamiliar vocabularies affect the item difficulty and item discrimination index (Pradanti et al., 2018).

3. RESEARCH METHODOLOGY

3.1 Research Design

This study used descriptive quantitative research since this study aims to find out the quality of test items of English final semester test for grade nine students in the academic year of 2018/2019 in Ponorogo. Anderson and Arsenault (2005) state that descriptive quantitative research aim to portray the data as a whole by grouping and representing the data in tables or figures.

3.2 Population and Sample

The population of this study was the grade nine students of 74 junior high schools. The researchers employed proportionate stratified random sampling to acquire the representative sample. The sample involved in this study were 151 samples in the form of students' answer sheets of English final semester test for grade nine students in the academic year 2018/2019 in Ponorogo. The students' answer sheets were from the different junior high schools which are already divided in three ranks: top, middle, and bottom rank.

3.3 Instruments

The researchers applied a blank table as an instrument in this study. The blank table refers to the Quest program report of multiple-choice test item analysis. Researchers used this table to record the calculation results of Quest program. This instrument involves 3 characteristics: item discrimination, item difficulty, and distractors in which these characteristics cover the quality of the test. An item was accepted when it conforms to the whole ideal index of item difficulty, item discrimination, and effectiveness of distractor. Conversely, an item was eliminated when it does not conform to one of the item difficulty, item discrimination, and effectiveness of distractor.

3.4 Data Analysis Procedures

To analyze the data, the researchers computed through the Quest program to obtain the calculation of item difficulty, item discrimination, and distractors index. The answer key and students' responses of the test package were typed in the form of notepad file. Afterward, researchers created file control in the form of notepad as a command to analyze the data. The file control must place in the same location with the Quest program. Then, the researchers ran the program and typed 'submit' word followed by the file control's name. Automatically, this program created output file which provided the calculation of item difficulty, item discrimination, and distractors index. The item difficulty index was presented in the form of a value percentage that has an asterisk symbol. The range index was from 0.00 to 1.00. For the discrimination index, the index was presented from biserial point value that has an asterisk symbol. The well discrimination index offered a positive index. Whereas for the distractors could be seen from the percentage of students who select the option. A distractor was effective when the biserial point value was lower than the biserial point of the correct option.

4. FINDINGS

There were 40 items in the form of multiple-choice test with 4 options in the English final semester test for grade nine students in the academic year of 2018/2019 in Ponorogo. Quantitative analysis was conducted to identify the quality of test items based on item difficulty, item discrimination, and distractors using Quest program. In general, the findings revealed that the index of item difficulty, item discrimination, and distractors is very high. The findings are determined with judgments: items were accepted when the items conformed to all of the three characteristics and items were eliminated when the items did not conform to one of the three characteristics.

4.1 Item Difficulty

The researchers calculate the item difficulty based on students' response. Table 1 displays the item difficulty index from Quest program related to the level difficulty of items.

Table: 1 Classification of Item Difficulty

Index	Category	Frequency	Percentage
0.00 - 0.30	Difficult	0 item	0%
0.31 - 0.70	Moderate	37 items	92.5%
0.71 - 1.00	Easy	3 items	7.5%

As Table 1 displays, the item difficulty values indicate that the 92.5% (37 items) involves in moderate category and 7.5% (3 items) involves in easy category. Surprisingly, none item involves in difficulty category. The results indicate that the test package is not an ideal test.

4.2 Item Discrimination

The next characteristics are showed in Table 2 which presents the calculation of Quest program related to item discrimination.

Table: 2 Classification of Item Discrimination

Index	Category	Frequency	Percentage
0.40 and up	Very good item	39 items	97.5%
0.30 - 0.39	Accepted item with little revision	1 item	2.5%
0.20 - 0.29	Need revision	0 item	0%
Below 0.19	Poor item, to be rejected	0 item	0%

The Quest program calculation reveals that the 97.5% (39 items) with very good discrimination index and 2.5% (1 item) needs little revision. Hence, most of the items include in accepted item and can be used as item bank. There is only an item should improve by little revision.

4.3 Effectiveness of Distractors

In the final characteristics, researchers analyze the effectiveness of distractors. The researchers focus on the biserial point of the options. The table below shows the distribution of distractors.

Table: 3 Distributions of Distractors

Category	Frequency	Percentage
Effective	32 items	80%
Ineffective	8 item	20%

As clearly presents in Table 3, most all of the distractors of English final semester test in the academic year of 2018/2019 in Ponorogo are effective to distract the students. What is interesting about the data from Quest program that there are 32 items (80%) as effective distractor and the other 8 items (20%) are ineffective distractor. The results indicate that most of items can distract students effectively.

5. DISCUSSION

This study set out with the aim of identifying the quality of the English final semester test in the academic year of 2018/2019 in Ponorogo based on item difficulty, item discrimination, and the effectiveness of distractor. In the current study, out of 40 test items, most of the items are acceptable in the item difficulty. The 37 items are moderate items while 3 items are easy items. The ideal test involves 25% easy items, 50% moderate items, and 25% difficult items (Kunandar, 2013). The results of this study do not show that the test package has proportional item difficulty. Alderson et al. (1995) said that this condition cannot reveal the exact students' ability. These results involve more moderate category than easy and difficult category. This argumentation confirmed Brown (2004) who argued a well-constructed item cannot be too easy or difficult. A test package should cover each difficulty level so that teachers can recognize the abilities of each student. By contrast, Haider et al. (2012) argued that the dominant category that is medium category could indicate that the students have well comprehension to answer the test since more than half of the students answer the items correctly. This can be related to none difficult items in the test package.

These results are comparable to those of other studies (e.g. Amelia, 2010; Maghfiroh, 2010; Ani, 2011; Risydah, 2014; Haryudin, 2015; Manfenrius et al., 2015; Pradanti et al., 2018; Maghfiroh, 2019), although test conditions do not similar. These previous studies reported the disproportionate portion among easy, moderate, and difficult items. There is possible explanation for these results. Item difficulty can be influenced by cognitive factors (Sung et al., 2015). Cognitive factors involve comprehension, coding, transition, scrutinizing, and working memory (Danili & Reid, 2006). They added that cognitive factors affect students' performance and achievement so these factors affect calculation of item difficulty.

These results are also likely to be related to factors which can affect item difficulty namely, theory of item analysis involves classical test theory (CTT) and item response theory (IRT), the clarity of items instruction, and the suitability between material and items (Haladyna, 2004). The use of statistical theory in analyzing the quality of items can affect the accuracy of the index results. Furthermore, the instruction of items also affects the students' comprehension which affects their answers automatically. Students might answer with incorrect answer when the questions contain unclear instructions. Last, the suitability of the materials with the questions also affects the item difficulty. Students would be difficult to answer the questions when the questions are not in accordance with the material that has been studied in class. In short, teachers or test makers should concern with these factors to achieve a balanced item difficulty.

On the question about item discrimination, the results are very great. The Quest program calculation revealed that the 97.5% (39 items) with excellent discrimination index and 2.5% (1 item) with poor discrimination index. These results indicate that 1 poor item needs revision. It is interesting to note that most of test items of this study can be kept as item bank and used for further test. These accords with Ebel and Frisbie (1991) that the great item discrimination index is influenced by the moderate item difficulty index.

As stated in the Quest program result, 92.5% of the test items are moderate items. The great item discrimination index leads the test items to discriminate high and low achievers.

These results differ from previous studies. To date, several previous studies (see Toha, 2010; Lestari, 2011; Risydah, 2014; Haryudin & Santosa, 2016; Pradanti et al., 2018; Maghfiroh, 2019) reported that the excellent items less than 50%. Other studies (e.g. Haryudin, 2015; Manfenrius et al., 2015; Fajriah, 2016) have reported zero excellent items. A possible explanation for this might be that homogeneity of options (Haladyna & Rodriguez, 2013). The options that are not homogenous in content and grammar cause students to find the right answer easier implicitly (Atalmis & Kingston, 2018).

Having defined item discrimination, the researcher will now move on to discuss the effectiveness of distractors. The effectiveness of distractors is recognized by analyzing distractors. The results of this study showed that there are 32 items (80%) which have effective distractors and 8 items (20%) which have ineffective distractors. The levels analyzed in this study are high above than the previous studies (see Risydah, 2014; Haryudin, 2015; Manfenrius et al., 2015; Pradanti et al., 2018; A. Maghfiroh, 2019). These previous studies found that more than 40% were ineffective distractors.

These results may be explained by the fact that the item discrimination index may have been an important factor in the effectiveness of distractors. As stated in the previous, most of the test items are able to discriminate between high and low achievers. It can therefore be assumed that the great item discrimination can lead to the effectiveness of distractors (Kheyami et al., 2018). Despite this, the ideal number of distractors affects the functionality of options. An item had at least three distractors to make the item work well (Haladyna, 2004; Rodriguez, 2005; Kheyami et al., 2018). Since this test package has three distractors in each test item, most of the items have effective distractors. Creating reasonable distractors and decreasing ineffective distractors were important to increase the test items' quality (Rodriguez, 2005) .

According to these results, we could infer that the test package was a good test. The 31 items (77.5%) were well-constructed item since they conformed to the characteristics of item analysis. While, the other 9 items (22.5%) were poor-constructed item since they do not conformed to the characteristics of item analysis. A good test was able to reveal the students' performance accurately (Quaigrain & Arhin, 2017). It indicated that there was suitability between the items and the material being studied (Gareis & Grant, 2015). A good test can build the effective and comfortable atmosphere classroom-learning because the teachers realize the students' needs and abilities. It automatically reveals the specific topics or materials which need more emphasis or clarity. Moreover, the students' higher level cognitive was able to assess (Quaigrain & Arhin, 2017). Mukherjee and Lahiri (2015) proposed that a well-constructed item is capable to assess higher level cognitive such as knowledge, application, analysis and synthesis. Furthermore, the effect of a good test made teachers easier to assess students' performance level and provided the consistent scores (Hotiu, 2006).

To achieve the consistent scores, improving the assessment literacy needs to be carried out by teachers and test makers because assessment is a complex, dynamic and continuous process. (Xu & Liu, 2009). The teachers and test makers who have a good assessment literacy are able to construct and apply tests with a high level of validity and reliability continuously (Gareis & Grant, 2015).

6. CONCLUSION

The Quest program results provide evidence that generally, the test is a good test. Although several items must be revised or replaced with the new item, most of the items conform to be well-constructed items. These poor items may be influenced by other causes such as, students' understanding level, ambiguity of instructions, difficult materials or topics, and ambiguity in the options or even key answer. In spite of this study has several advantages, it contains several limitation such as the few variable and data of the study. Access to offer seminars on constructing test item must to be found. These results may support teachers or test makers as an effective feedback to change in the way they construct test items. Moreover, the way teachers teach and the atmosphere of teaching-learning activity can be improved. In the future study, other researchers should add other techniques of analyzing test item to compare the results. Other researchers should also complete the study by qualitative analysis to obtain the deeper findings. The students' argumentation may be included for discovering more accurate about the level of difficultness items and enhancing the solution of the problems.

7. REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge University Press. http://en.bookfi.net/book/1052551

Amelia, R. (2010). *An Analysis of the English Summative Test Items in Terms of Difficulty Level* [Bachelor Thesis, Universitas Islam Negeri Syarif Hidayatullah Jakarta]. http://repository.uinjkt.ac.id/dspace/handle/123456789/5888

Anderson, G., & Arsenault, N. (2005). *Fundamental of Educational Research* (2nd ed.). Taylor & Francis e-Library. http://en.bookfi.net/book/1145351

Ani, L. A. (2011). *An Item Analysis on the Difficulty Level of an English Summative Test for Second Grade of SMP Muhammadiyah 29 Cinangka-Sawangan Depok* [Bachelor Thesis, Universitas Islam Negeri Syarif Hidayatullah Jakarta]. http://repository.uinjkt.ac.id/dspace/handle/123456789/4878

Arikunto, S. (2016). *Dasar-Dasar Evaluasi Pendidikan* (2nd ed.). Bumi Aksara.

Atalmis, E. H., & Kingston, N. M. (2018). The Impact of Homogeneity of Answer Choices on Item Difficulty and Discrimination. *SAGE Open*, *8*(1), 1–9. https://doi.org/10.1177/2158244018758147

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*, *2*(2), 189–193.

Browder, D. M., Wakeman, S., & Flowers, C. P. (2006). Assessment of Progress in The General Curriculum For Students With Disabilities. *Theory into Practice*, *45*(3), 249–259. https://doi.org/10.1207/s15430421tip4503_7

Brown, H. D. (2001). *Teaching by Principles: An Interactive Approach to Language Pedagogy* (2nd ed.). Longman.

Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. Longman.

Danili, E., & Reid, N. (2006). Cognitive Factors That Can Potentially Affect Pupils' Test Performance. *Chemistry Education Research and Practice*, *7*(2), 64–83.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5th ed.). Prentice-Hall. http://en.bookfi.net/book/1103329

Fajriah, S. (2016). *An Item Analysis on Discriminating Power of English Summative Test* [Bachelor Thesis]. Universitas Islam Negeri Syarif Hidayatullah Jakarta.

Flucher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advance Resource Book*. Routledge.

Gareis, C. R., & Grant, L. W. (2015). *Teacher-Made Assessments: How to Connect Curriculum, Instruction, and Student Learning* (2nd ed.). Routledge.

Haider, Z., Latif, F., & Mushtaq, M. (2012). Evaluation of English Achievement Test: A Comparison Between High and Low Achievers Amongst Selected Elementary School Students of Pakistan. *Educational Research and Reviews*, *7*(29), 642–650. https://doi.org/10.5897/ERR12.035

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items* (3rd ed.). Lawrence Erlbaum Associates Publisher.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. Routledge. https://b-ok.asia/book/2329518/e04c25

Haryudin, A. (2015). Validity and Reliability of English Summative Tests at Junior High School in West Bandung. *Jurnal Ilmiah UPT P2M STKIP Siliwangi*, *2*(1), 77–90.

Haryudin, A., & Santosa, I. (2016). The Analysis of Discriminating Power of English Summative Test. *Jurnal Ilmiah UPT P2M STKIP Siliwangi*, *3*(2), 59–67.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of One-Best MCQs: The Difficulty Index, Discrimination Index and Distractor Efficiency. *JPMA-Journal of the Pakistan Medical Association*, *62*(2), 142–147.

Hotiu, A. (2006). *The Relationship Between Item Difficulty and Discrimination Indices in Multiple-Choice Tests in A Physical Science Course* [Master Thesis, Florida Atlantic University]. http://www.physics.fau.edu/research/education/A.Hotiu_thesis.pdf

Izard, J. (2005). *Trial Testing and Item Analysis in Test Construction*. UNESCO International Institute for Educational Planning. http://www.sacmeq.org

Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item Analysis of Multiple Choice Questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos Univesity Medical Journal*, *18*(1). https://doi.org/10.18295/squmj.2018.18.01.011

Kunandar, K. (2013). *Penilaian Autentik: (Penilaian Hasil Belajar Peserta Didik Kurikulum 2013)*. RajaGrafindo Persada.

Lestari, H. (2011). *An Item Analysis on Discriminating Power of English Summative Test* [Bachelor Thesis]. Universitas Islam Negeri Syarif Hidayatullah Jakarta.

Maghfiroh, A. (2019). *An Analysis on English Midterm Test for the Second Semester at the Ninth Grade of SMP TA'MIRUL ISLAM Surakarta* [Bachelor Thesis]. Institut Agama Islam Negeri Surakarta.

Maghfiroh, F. (2010). *An Item Analysis of the Difficulty Level of an English Summative Test* [Bachelor Thesis, Universitas Islam Negeri Syarif Hidayatullah Jakarta]. http://repository.uinjkt.ac.id/dspace/bitstream/123456789/3860/1/FIFI%20MAGHFIROH-FITK.pdf

Manfenrius, A., Sutapa, G., & Wijaya, B. (2015). Item Analysis on English Summative Test at The Eighth Grade Junior High Schools in Pontianak. *Jurnal Pendidikan Dan Pembelajaran Khatulistiwa*, *4*(12), 1–10.

Mardapi, D. (2015). *Pengukuran, Penilaian, dan Evaluasi Pendidikan*. Nuha Litera.

Peraturan Menteri Pendidikan dan Kebudayaan RI tentang Standar Penilaian Pendidikan, Pub. L. No. 23 (2016). http://arxiv.org/abs/1011.1669

Miller, D. M., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and Assessment in Teaching* (10th ed.). Pearson Education.

Mukherjee, P., & Lahiri, S. K. (2015). Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal. *IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)*, *14*(12), 47–52.

Ofianto, O. (2018). Analysis of Instrument Test of Historical Thinking Skills in Senior High School History Learning with Quest Programs. *Indonesian Journal of History Education*, *6*(2), 184–192.

Pradanti, S. I., Martono, M., & Sarosa, T. (2018). An Item Analysis of English Summative Test For The First Semester of The Third Grade Junior High School Students in Surakarta. *English Education*, *6*(3), 312–318. https://doi.org/10.20961/eed.v6i3.35891

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*, 1–11.

Risydah, Y. (2014). *An Analysis of Test Items of the English First-Term Test of the Seventh Grade Students of SMP Muhammadiyah 10 Yogyakarta in the Academic Year of 2013/2014* [Bachelor Thesis, Universitas Negeri Yogyakarta]. https://eprints.uny.ac.id/18498/1/Yunita%20Risydah%2006202244051.pdf

Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice*, 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Rosana, D., & Setyawarno, D. (2017). *Statistik Terapan Untuk Penelitian Pendidikan*. UNY Press.

Scouller, K. (1998). The Influence of Assessment Method on Students' Learning Approaches: Multiple Choice Question Examination Versus Assignment Essay. *Higher Education*, *35*(4), 453–472. https://doi.org/10.1023/A:1003196224280

Setiyana, R. (2016). Analysis of Summative Tests for English. *English Education Journal*, *7*(4), 433–447.

Sulistyo, G. H., & Suharyadi, S. (2018). The Profile of EFL Learners As Measured by An English Proficiency Test. *JEELS (Journal of English Education and Linguistics Studies)*, *5*(1), 115–145. https://doi.org/10.30762/jeels.v5i1.570

Sung, P. J., Lin, S. W., & Hung, P. H. (2015). Factors Affecting Item Difficulty in English Listening Comprehension Tests. *Universal Journal of Educational Research*, *3*(7), 451–459. https://doi.org/10.13189/ujer.2015.030704

Suyata, P. (2016). *Program QUEST - Salah Satu Cara Meningkatkan Validitas Internal Penelitian Bahasa Indonesia*. 112–111.

Toha, H. (2010). *An Item Analysis of English Summative Test for the First Year of Junior High School* [Bachelor Thesis]. Universitas Islam Negeri Syarif Hidayatullah Jakarta.

Xu, Y., & Liu, Y. (2009). Teacher Assessment Knowledge and Practice: A Narrative Inquiry of a Chinese College EFL Teacher's Experience. *TESOL QUARTERLY*, *43*(3), 493–513.