www.bircu-journal.com/index.php/birci
Budapest International Research and Critics Institute-Journal (BIRCI-Journal)
Humanities and Social Sciences
ISSN 2615-3076 Online)
ISSN 2615-1715 (Print)
Budapest Institute

Developing Item Analysis of Teacher–Made Test for Summative Assessment of Seventh Grade of SMPN 8 Komodo in Academic Year 2020/2021

Aloysia Moto[1], Lailatul Musyarofah[2], Kani Sulam Taufik[3]
[1,2,3]*STKIP PGRI Sidoarjo, Indonesia*
aloysiamoto@gmail.com, Ibulaila7810@gmail.com, sullamtaufik@gmail.com

Abstract

It was the goal of this study to improve the analysis of teacher-made English test items and to evaluate their quality. The researchers used the Research & Development (R&D) procedure to adapt the ADDIE model. @Aloytemsis, the development's product, became the answer as a result of a needs assessment. At SMPN 8 Komodo, the validity, reliability, index difficulty, and discriminating power of the summative test items were all graded. 43 students who had to take the summative test participated in the field testing. In case of validity, it indicated that 15 items (37.5%) were classified into 'invalid' due to the validity index were < 0.279. Meanwhile, 25 items (62.5%) were classified into 'valid' criteria regarding the value of validity index were > 0.279. Then, the items under consideration received a reliability index of 0.762. In other words, the "teacher-made English summative test items" were deemed "Good Reliable". It was discovered that the tough category contained 22 items (55%), the medium category contained 18 items (45%), and the easy category contained just 0 items (0%). Lastly, it is found that the items test were 13 items (32.5%) were in the category 'Poor' items due to its value of discrimination index 0.00-0.19. Then, in the category 'Enough', there were 12 items (30%). On the other hand, there were 15 items (37.5%) were classified into 'Good' items. Unfortunately, there was no items into category 'Very Good' items.

Keywords
item analysis; teacher-made test; summative assessment

I. Introduction

In the development of the world of education, especially after the rolling reforms, new phenomena have arisen in educational institutions, which are schools that use the term Integrated Islamic Schools (Titik, 2010: 42). The school is essentially aimed at helping parents teach good habits and add good character, also given education for life in society that is difficult given at home. Thus, education in schools is actually part of education in the family, which is also a continuation of education in the family (Daulay in Ayuningsih, W. et al. 2020).

Testing and teaching are inextricably linked because the effectiveness of a teacher's instruction cannot be measured and determined without first administering a test. The test is designed to determine the level of achievement achieved by students during the teaching and learning process. Then, based on the results, it can be determined whether the learning objectives have been achieved successfully or not, allowing for the identification of students' weaknesses and the making of an informed decision. It is dependent on the types of tests that are performed. A test, according to Brown, is designed to assess a person's ability, knowledge, and performance in a given situation (Brown, 2003). Additionally, a test may be designed to assess students' readiness to enter the program or to determine whether or not they are prepared to continue in the program after they have entered it (Bachman, 1990).

DOI: *https://doi.org/10.33258/birci.v5i1.4237*

According to Bloom (1971), testing or evaluation has a broad perspective and a specific place in education such as (a) evaluating a student's level of learning and teaching effectiveness by gathering and analyzing evidence; (b) examining a wide range of evidence instead of just a final paper and pencil test; (c) evaluation as a process for determining the degree to which students are developing in the desired manner and for clarifying the most important educational goals and objectives; and (d) to determine whether or not a teaching-learning process is effective and, if not, what changes need to be made to ensure its effectiveness before it is too late, evaluation serves as a feedback-corrective system. A comprehensive test of educational research and practice to determine whether or not alternative methods are just as effective in achieving specific educational goals. It's been said that it's impossible to evaluate student progress without administering tests. The ability of a student to assess and learn is assessed through a test. The value of a good exam or criterion must also be understood by teachers in order to value school evaluations (Arikunto, 2005). The learning objectives and goals of students can also be tested. As a result, tests must be well-structured and comprehensive.

Tests, then, are a means of gauging how well students have learned during the teaching-learning process. As a consequence, teachers should be able to plan and conduct an effective assessment. As a result, teachers' ability to accurately and carefully assess students' abilities may have a significant impact on the improvement of teaching quality. This information is beneficial to both students and teachers in their educational endeavors. Students and teachers alike can benefit from this tool, which serves as both a means of gauging progress and a means of providing feedback. In light of the significance of the evaluation, it is necessary to keep in mind that a good test is one that is well constructed. A good test should fulfill specific the criteria. According to some experts, there are four criteria of a good test: validity, reliability, level of difficulty, and discrimination power (Bichi, 2015). A reliable and valid test requires a methodical selection of test items. For an exam to be reliable, it must be graded consistently across all students. Each item has an impact on the test's quality. The item analysis follows the administration and scoring of the test on the chosen samples.

There are two ways to conduct a test. Both a teacher-created and a standard-bearer-created test exist (Brown, 2003). A teacher-made test is one that the teacher has designed themselves. An assessment based on what was covered in class and how it was implemented is common practice for teachers. In this case, the teacher has the option of taking the test (Widodo, & Slamet, 2021). The purpose of a teacher-made test is to assess whether or not students have learned enough to meet the curriculum's objectives. Because of this, educators must pose questions that are grounded in logic and reason. This test is usually used for daily, formative, and general tests (summative). In this research, the summative test was considered as a main discussion due to it was applied when this research being conducted.

Today, most English language tests at a junior high school are made by the teachers at that institution. This is one of the impacts of implementing the newest curriculum (The 2013 curriculum). The 2013 curriculum applies authentic assessment to assess student learning progress. Legitimate processes, materials, and tests are inseparably linked to valid assessment. Bloom's taxonomy serves as the primary guide for the implementation of the 2013 curriculum. Teachers can use Bloom's Taxonomy to help them analyze and align their lesson plans, instructional methods, and classroom assessment. As defined by Bloom's taxonomy, educational goals are classified as either affective, psychomotor, or cognitive in Bloom's classification system. Bloom's taxonomy, like other taxonomies, is hierarchical, which means that learning at the higher levels requires prior knowledge and

skills. Motivating teachers to cover all three areas of Bloom's Taxonomy is one of its primary objectives. Bloom identified six cognitive levels, starting with the lowest level of simple recall or recognition of facts and progressing to the highest level of complex and abstract mental processes. The classifications are as follows: Remembering and Understanding categorized as LOTS (Lower Order Thinking Skills), Applying categorized as MOTS and the last are Analyzing, Evaluating and Creating which called as HOTS (Higher Order Thinking Skills). Bloom's Taxonomy classification can be used to determine the type of difficulty level. Based on it the English teacher has to be able to construct their good test.

The majority of teacher-created tests are in the form of multiple-choice questions. The construction of multiple-choice items is difficult and time consuming, but scoring them is simple. It is necessary to attempt an items analysis in order to produce or construct a good test, particularly a multiple-choice test. In Hughes (1989), the stages of test construction are divided into three main stages. Before pretesting, the writer can begin by writing the test specifications and then proceed to writing the test itself. When it comes to constructing test items, however, teachers today lack the necessary skills and techniques. According to Hughes (1976), a lot of high-quality language testing fails to measure what it is supposed to measure. In some cases, a test designed by a teacher may only measure a student's ability, rather than the teacher's goals for the test. The teacher-created test lacks validity and reliability when compared to the school's academic standards.

In order to gain a good test, the item analysis is needed to be provided regarding it is a process for analyzing the student responses to the different test items in order to assess the quality and quality of the test as a whole, enabling teachers to increase their testing skills, identify specific fields of content which need to be emphasized or clarified and improve other classroom practices (Bichi, 2015, p. 1655). Item analyses will be especially valued in developing items that will again be used in later tests, but can also be used in a single test administration to eliminate ambiguous or misleading items. Furthermore, item analysis is valuable for enhanced instructors in test design and the identification of specific areas of content that require greater emphasis or clarity. For each raw score, separate item analysis may be requested. In addition, the test under analysis collects items that measure a single subject area or underlying skill. The quality of the test in its entirety is assessed by evaluating its 'internal consistencies' and by comparing the item responses of the students with their overall test results (Fatimah, Elzamzami, & Slamet, 2020).

Regarding the researcher's observation at the summative test of seventh grade of Junior High School, the researchers found many test-made by the English teacher were copied from the Internet, the way the teacher makes the test and does not consider the syllabus and curriculum. The test items only reflect the understanding of the fact of the lesson and perceptions of the goals of education which all educators do not share. Concerning the explanations mentioned, the researchers were interested in analyzing the English multiple-choice items test as a summative assessment of the 7th grade students at SMPN 8 Komodo in the academic year 2020/2021.

This research was conducted to reveal the test score's reliability, discrimination power, index difficulty, and item analysis to provide detailed information leading to test item construction improvement. Therefore, hopefully, the findings of this research will provide a deeper understanding and important information for the teachers and other researchers in regard that that analyzing items test is part of continuing professional development for teachers. Based on the previous explanation, the main problems of this research were: (1) How is the product of Items Analysis of Teacher- Made Test for seventh

grade's Summative Assessment of SMPN 8 Komodo designed and (2) How is the quality of test items for seventh grade's Summative Assessment constructed by the Teacher in SMPN 8 Komodo?

II. Research Method

The research design of this study belongs to Research and Development, or commonly abbreviated as R & D. Research and development can be defined as a process or steps to develop a new product or to complete an available product and be accountable (Sujadi, 2003:164). According to Sugiono (2010:407), research and development is a research method which is used to produce a certain product and examine the effectiveness of the product. Borg and Gall (1983:772) define the research and development method as a process which is used to develop and validate educational products.

Based on these opinions, the research is adapting the ADDIE model. The researchers chose the ADDIE model as this research model because the ADDIE model is quite simple and appropriate for educational research and development. ADDIE model consists of analysis, design, development, implementation, and evaluation. This research aimed to obtain systematic information related to the quality of the tests. Qualitative analysis is fulfilled by using a format of validation sheet or expert review. It is distributed to identify the quality of the tests based on the principles of writing a test. The principles are the appropriateness of the material, construction, and language. Then, the results included the validity, reliability, level of difficulty, discrimination index, and distractors power. This study was conducted at SMPN 8 Komodo. The object involved in this study was the assessment test as a final test of the 7th-grade student in the academic year 2020/2021 that was made by the English teachers and the students' answer sheets. The final test consists of 40 items of multiple choice. The researcher took 20 answer sheets from 43 students answer sheets based on random sampling at least 10% from the population (Arikunto, 2010).

III. Discussion

A development product in the form of items analysis was created as a result of this research, which was given the name @Aloytemsis. The product was developed with the help of Microsoft Excel as a main basis. It was determined that the product's content was aligned with the objectives of the teacher-made summative English test, and it was developed in accordance with the findings of the need analysis.

3.1 Product Development

Obtaining information from the administrator of SMPN 8 Komodo as well as the Head of the Master of SMPN 8 Komodo in relation to the summative test that had been completed previously served as the basis for the need analysis. In order to respond to the findings of the needs analysis, a product development plan for the items analysis was put together. Following the results of the need analysis, it was discovered that the development of content related to the items of the teacher-made test in SMPN 8 Komodo was required due to the fact that the test's items were of questionable quality in terms of their overall design. Additional manual scoring of the tests was carried out by the test creators and team in order to ensure that the results of the tests were accurate. Because of this, developing a product from the items analysis could be a solution to the problem that will allow it to be resolved in the long run. @Aloytemsis is the name of the final product of the development.

The product development and analysis of the items were carried out using a Microsoft Excel database. On the front page, there are several buttons that can be used to navigate around the site and conduct the item analysis. The first line contains three buttons, namely 'Item Analysis' as the menu page's title, 'Input Data 1' for navigating the test's data related to the 'Test Results' section, 'Input Data 2' for computing the test's data related to the 'Analyzing' section, and 'Input Data 3' for inputting the test's data related to the 'Total Score' section.

On the sheet 'Score', the results of the test takers' scores were displayed. This sheet contains data pertaining to the test's results. The top area contains the identity of the SMPN 8 Komodo institution, as well as information about the test that was conducted. The primary section contains information about the results of each section, such as 'Section 1', 'Section 2', and 'Section 3', as well as the converted scores based on the provided rubrics. Additionally, this sheet displays the test takers' ranks in relation to their grade.

3.2 Items Analysis

The need analysis was conducted by obtaining information from the administrator at SMPN 8 Komodo and the Head of the Master at SMPN 8 Komodo regarding the previously completed summative test. The product development of the items under analysis was organized around the test results concerning validity, reliability, index difficulty and discriminating power.

a. Validity

The validity of the formula correlation point biserial was determined as part of the product development process using the @Aloytemsis formula. It was determined that all of the students served as the sample for the product's field testing. The accompanying table indicates that 30 items were classified as valid during the product's field testing in the test section. As illustrated in the accompanying table, the item's validity analysis yielded satisfactory results.

Table 1. The Validity of the Items

No	Validity Index	Item Question	Amount	Percentage
1	< 0.279 (invalid)	3, 5, 6, 9, 11, 12, 14, 18, 20, 24, 26, 28, 31, 32, 39, 40	15	37.5%
2	$\geq$ 0.279 (valid)	1, 2, 4, 7, 8, 10, 13, 15, 16, 17, 19, 21, 22, 23, 25, 27, 29, 30, 33, 34, 35, 36, 37, 38	25	62.5%

The table indicated that 15 items (37.5%) were classified into ‘invalid’ due to the validity index were < 0.279. those items were 3, 5, 6, 9, 11, 12, 14, 18, 20, 24, 26, 28, 31, 32, 39, and 40. Meanwhile, 25 items (62.5%) were classified into ‘valid’ criteria regarding the value of validity index were $\geq$ 0.279. those items were 1, 2, 4, 7, 8, 10, 13, 15, 16, 17, 19, 21, 22, 23, 25, 27, 29, 30, 33, 34, 35, 36, 37, and 38.

b. Reliability

After conducting a @Aloytemsis, the reliability of the items test results was compared to a criterion that states that when the correlation coefficient is less than 0.312, previously considered trustworthy items are now considered less trustworthy. If a question's reliability coefficient (r_{11}) is less than 0.312, it is possible that it was previously unreliable or had a low level of reliability. The items under consideration received a reliability index of 0.762 during the examination of the test scores, which was used to

determine the section's overall trustworthiness. This means that the “teacher-made English summative test items” were regarded to have *“Good Reliability”* based on the results.

c. Index Difficulty

For the purposes of interpretation, classification was used to group questions into categories. The results of the calculation of the level of difficulty were divided as follows: 0,00-0,29 for category difficulty questions; 0,30-0,69 for category medium questions; 0,70-1,00 for category easy questions. After conducting an investigation through @Aloytemsis it was discovered that the tough category contained 22 items (55%), the medium category contained 18 items (45%), and the easy category contained just 0 items (0%) based on the results of the investigation. The following is a list of the difficulty levels in items analysis:

Table 2. The Distribution of Difficulty Index of the Items

No	Difficulty Index	Item Question	Amount	Percentage
1	0,00 – 0,29 (difficulty)	1, 4, 6, 7, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 27, 30, 34, 38	18	45%
2	0,30 – 0,69 (medium)	2, 3, 5, 8, 10, 11, 12, 24, 26, 28, 29, 31, 32, 33, 35, 36, 37, 39, 40	22	55%
3	0,70 – 1,00 (easy)	-	0	0%

According to the table, 18 items (45%) were classified as ‘difficult' because their difficulty index was 0.00-0.29. The items were 1, 4, 6, 7, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 27, 30, 34, and 38. Meanwhile, 22 items (55%) met the ‘medium' criteria when the value of the difficulty index was 0.30-0.69. These were the following numbers: 2, 3, 5, 8, 10, 11, 12, 24, 26, 28, 29, 31, 32, 33, 35, 36, 37, 39, and 40. Fortunately, there was no items were classified into ‘easy’.

d. Discriminating Power

For the purposes of interpretation, classification was used to group questions into categories. A categorization was used to interpret the results of the calculation of the ‘Discrimination Power’, with values ranging from 0,00 to 0,19 falling into the category of 'Poor'; 0,20 to 0,39 falling into the category of 'Enough'; 0,40-0,69 falling into the category of 'Good,' and 0,70-1,00 falling into the category of 'Very Good.' Following the results of the analysis undertaken by @Aloytemsis, it was discovered that items of the teacher-made English summative test at SMPN 8 Komodo scored in the 'Poor' category of the 'Discrimination Index' totaled 13 items (32.5%), 12 items (30%), 15 items (37.5%), and 0 item (0%) went on to score in the 'Very Good' category of the 'Discrimination Index'. On the basis of the discrimination index, the following 40 items were presented:

Table 3. The Distribution of Discrimination Index of the Items

No	Discrimination Index	Item Question	Amount	Percentage
1	0,00-0,19 (Poor)	13, 16, 20, 21, 23, 24, 25, 27, 29, 30, 31, 36, 38	13	32.5%

2	0,20-0,39 (Enough)	4, 5, 15, 16, 20, 22, 28, 32, 33, 34, 39, 40	12	30%
3	0,40-0,69 (Good)	1, 2, 3, 6, 7, 8, 9, 10, 11, 14, 17, 18, 26, 35, 37	15	37.5%
4	0,70-1,00 (Very Good)	-	0	0%

Concerning to the table above, it is indicated that the items of the teacher-made English summative test at SMPN 8 Komodo were 13 items (32.5%) were in the category 'Poor' items due to its value of discrimination index 0.00-0.19. those items were 13, 16, 20, 21, 23, 24, 25, 27, 29, 30, 31, 36, and 38. Then, in the category 'Enough', there were 12 items (30%) which were 4, 5, 15, 16, 20, 22, 28, 32, 33, 34, 39, and 40. On the other hand, there were 15 items (37.5%) were classified into 'Good' items. Those items were 1, 2, 3, 6, 7, 8, 9, 10, 11, 14, 17, 18, 26, 35, and 37. Unfortunately, there was no items into category 'Very Good' items.

In response to the needs analysis findings, a product development plan for the items analysis was created. The need analysis revealed that teacher-created English tests or assessments should be reviewed in advance due to the item quality. Additionally, the creators of the tests and their team manually scored the tests to confirm their findings. As a result, developing a product based on the item analysis may be a feasible option. @Aloytemsis is currently being developed to address the issue. By utilizing item analysis to identify areas of course content that require students' attention or clarification, educators can create examinations and improve their ability to assess students' comprehension of those areas of topic. This means that the exam's "internal consistency" was used to determine its overall quality (Putri, 2015, p. 31).

When developing a good test, it is essential to consider the validity and reliability of the results because they are of critical importance (Brown, 2004, p. 19). A sheet area for "Items Analysis" was included in the part on @Aloytemsis after researchers were able to ascertain the quality of the item analysis. This sheet area contains data about the test's validity and reliability as well as information about index difficulty and index discrimination. The extent to which the established criteria for evaluating test results are met is referred to as this type of validity evidence. According to established criteria such as commercially produced tests or observed behavior, the results of this assessment are then compared to the results of a previous assessment (Fulcher & Davidson, 2007 & Brown, 2004). Furthermore, the best single measure of test accuracy is the degree to which test results are consistent, stable, and error-free. A test's reliability can be gauged by looking at how similar the results are between administrations, which is known as the Coefficient Alpha (or KR-20) (McCowan & McCowan, 1999, p. 10). Validity of a reliable test is not guaranteed. The degree to which an instrument is consistent is known as its reliability. According to the criteria set, what can be trusted about a scrupulous and test? This is a reliability test. A test is considered reliable if it consistently yields the same results when applied to the same group at various points in time (Zainal Arifin, 2011, p. 258).

In this study, the results of the reliability test were compared to a criterion that states that if r_{11} is less than 0.312, previously trusted items are now less trustworthy. For questions with reliability coefficients (r_{11}) less than 0.312, it is possible that the question has previously been unreliable or inconsistent. The items on the teacher-created test had a reliability index of 0.762, which was used to determine the overall reliability of the section as a whole. According to the items examined, the test items component at SMPN 8 Komodo demonstrated "Good Reliability" (DIIA, 2003).

The following is a breakdown of the difficulty rating results: 0,00-0,29 for difficult questions; 0,30-0,69 for medium questions; and 0,70-1,00 for easy questions fall into this range. Investigations conducted through @Aloytemsis revealed that the tough category had 22 items (55%), medium had 18 (45%), and easy had zero items (0%) based on the results of the investigations. The results of the 'Discrimination Power' calculation were categorized, with values ranging from 0,00 to 0,19 falling into the 'Poor' category; 0,20 to 0,39 falling into the 'Enough' category; 0,40-0,69 falling into the 'Good' category; and 0,70-1,00 falling into the 'Very Good' category. This categorization was used to interpret the results. According to @Aloytemsis' analysis of the teacher-made English summative test at SMPN 8 Komodo, 13 items (32.5 percent), 12 items (30%), and 15 items (37.5%) were found to be in the 'Poor' category of the 'Discrimination Index', while 0 items (0%) were found to be good. The difficulty index of the tests' items was found to be high. Discrimination increases as the value of the item increases. High-scoring students were more likely to correctly identify the item, whereas lower-scoring students were more likely to incorrectly identify it (DIIA, 2003; Jan Patock, 2004).

To ensure that the teaching materials developed by the researcher are appropriate for use, they must be reviewed by experts. According to the experts' statement, the materials for teaching were thoroughly reviewed and re-evaluated to make sure that the experts' statement was followed. From the observation sheets that had been given to experts, the results of the evaluation have been quoted in full. On the observation sheet, in addition to the material's systematic content, there was also a section for 'A Good Mark' drawings that had already been validated by a 'A Good Mark' on the observation sheet. Once all the items were marked and the results were positive, the product development team at @Aloytemsis indicated 'Valid.' In line with the advice of experts, we took the following action: For the reasons listed below, this is the case: The materials have been meticulously formulated. On the basis of this test, the purpose and goals of the project have been established, and a development framework has been established for the concept of integrated skills.

IV. Conclusion

A Microsoft Excel database was used as a starting point for product development for the items analysis, and it was then modified as needed that could be seen on the front-page navigation buttons to operate the items analysis, which can be used to operate the item analysis. As part of the product development process, the @Aloytemsis formula was used to determine the validity of the formula correlation point biserial. Every single student was used as a test group in order to conduct field testing of this new product. Using the accompanying table, we can see that 30 of the product's field-testing items were deemed valid. @Aloytemsis results were compared to criteria that state that when the correlation coefficient is less than 0.3112, previously trusted items are now considered to be less trustworthy. An unreliable or lowly reliable question's reliability coefficient is less than 0.312 if the reliability coefficient (r11) is less than 0.3112. As a result of the examination of test scores, a reliability index of 0.762 was used to determine the section's reliability. In other words, the "teacher-made English summative test items" were deemed to have “Good Reliability”.

The following is a breakdown of the difficulty rating results: 0, 00-0, 29 for difficult questions; 0,30-0,69 for medium questions; and 0,70-1,00 for easy questions fall into this range. Investigations conducted through @Aloytemsis revealed that the tough category had 22 items (55%), medium had 18 (45%), and easy had zero items (0%) based on the results of the investigations. The results of the 'Discrimination Power' calculation were

categorized, with values ranging from 0,00 to 0,19 falling into the 'Poor' category; 0,20 to 0,39 falling into the 'Enough' category; 0,40-0,69 falling into the 'Good' category; and 0,70-1,00 falling into the 'Very Good' category. This categorization was used to interpret the results. After analyzing the results of the teacher-made English summative test at SMPN 8 Komodo, it was discovered that 13 items (32.5%), 12 items (30%), 15 items (37.5%), and 0 items (0%) were found to be in the 'Poor' category of the 'Discrimination Index'.

Future studies on the analysis of placement test items or other topics associated with current research can benefit from this study's findings, according to the researchers. Item analysis of a test designed to resemble the tests was the focus of this research. This category of good test evaluated a test's validity, reliability, index difficulty, and discrimination index. This investigation's scope was constrained by the use of Microsoft Excel. A product suite that includes an online test, such as one that is integrated into a Learning Management System (LMS) or one that is integrated via the Internet, can be developed using the Research and Development (R&D) method, which is encouraged for researchers in the future.

References

Amari. (1997). Penilaian Pencapaian Hasil Belajar Mengajar Siswa di Sekolah.

Arikunto, Suharsimi. (2003). Dasar-dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.

Ayuningsih, W. et al. (2020). Implementation of Islamic Education Curriculum Development in Al-Ulum Islamic School Medan. Budapest International Research and Critics in Linguistics and Education (BirLE) Journal. P. 1033-1044.

Beacham, Lyle F. (1990). Fundamental Consideration in Language Testing. New York: Oxford University Press.

Bichi, A. A. (2015). Item Analysis using a Derived Science Achievement Test Data. International Journal of Science and Research (IJSR) Volume 4 Issue 5, May 2015, 1655-1662.

Bloom, S. Benyamin. George Madaus F. Thomas Hasting J. (1971). Evaluation to Improve Learning. New York: cc.

Brown, H.D. (2003). Language Assessment Principle and Classroom Practices. California: Longman.

Fatimah, S., Elzamzami, A. B., & Slamet, J. (2020). Item Analysis of Final Test for the 9th Grade Students of SMPN 44 Surabaya in the Academic Year of 2019/2020. *JournEEL (Journal of English Education and Literature)*, *2*(1), 34-46.

Fraenkel, R. Jack and Norman. (1993). How to Design and Evaluate Research in Education. Singapore. McGraw-Hill, Inc.

Heaton, J.B. (1975). Writing English Language Test. New York: Longman.

Heaton, J.B. (1990). Classroom Testing, New York: Longman Inc.,

Heaton, J.B. (1998). Writing English Language test. London and New York: Longman.

Hughes, Arthur. (2003). Testing for Language Teachers, New York: Cambridge University Press.

Johnson, David W. and Johnson, Roger T. (2002). Meaningful Assessment. A Manageable and Cooperative Process. USA: Allyn and Bacon.

Kemdikbud, (2017). Panduan Pembelajaran Untuk Sekolah Menengah Pertama /Madrasah Tsanawiyah (Smp/Mts). Jakarta,

Keputusan Kepala Badan Penelitian Dan Pengembangan Dan Perbukuan Nomor 018/H/Kr/2020 Tentang Kompetensi Inti Dan Kompetensi Dasar Pelajaran Pada

Kurikulum 2013 Pada Pendidikan Anak Usia Dini, Pendidikan Dasar, Dan Pendidikan Menengah Berbentuk Sekolah Menengah Atas Untuk Kondisi Khusus, 2021

Nurgiyantoro, Burman. (1995). Penelitian Dalam Pengajaran Bahasa dan Sastra. Yogyakarta: BPFE.

Sabat, Y., & Slamet, J. (2019). Studentsâ€™ Perception towards Written Feedback of Thesis Writing Advisory at STKIP Sidoarjo. *JET ADI BUANA*, *4*(1), 63-79.

Slamet, J., & Sulistyaningsih, S. (2021). Students' Difficulties in Answering "Structure and Written Expression" TOEFL-like at STKIP PGRI Sidoarjo. *E-Structural (English Studies on Translation, Culture, Literature, and Linguistics)*, *4*(01), 17-27.

Slamet, J., Sabat, Y., & Prasetyo, Y. (2019). *Students' Perceptions Toward Lecturers' Written Feedback Of Thesis Writing Advisory On The 7th Semester Students At STKIP PGRI Sidoarjo* (Doctoral Dissertation, Stkip Pgri Sidoarjo).

Weir, Cyrill. (1993). Understanding and Developing Language Tests. UK: Prentice-Hall International Ltd.

Widodo, J. P., & Slamet, J. (2021, December). Lecturers' Perspectives Through E-learning by Using Moodle for Post-Graduate Students at STKIP PGRI Sidoarjo. In *International Seminar on Language, Education, and Culture (ISoLEC 2021)* (pp. 167-171). Atlantis Press.

Widodo, J. P., & Slamet, J. (2020). Students'perception Towards Google Classroom As E-Learning Tool (A Case Study of Master of English Education of the Second Semester at STKIP PGRI Sidoarjo). *Magister Scientiae*, *2*(48), 99-109.