Journal of English Education and Teaching (JEET)
e-ISSN: 2622-5867
p-ISSN: 2685-743x
Volume 6 number 4, 2022
Page 498-513

A Test-Items Analysis of English Teacher-Made Test

Muslim Darmawan
Universitas Tanjungpura
muslimdrmwn@gmail.com
Sudarsono
Universitas Tanjungpura
sudarsono@fkip.untan.ac.id
Dwi Riyanti
Universitas Tanjungpura
dwi.riyanti@fkip.untan.ac.id
Yohanes Gatot Sutapa Yuliana
Universitas Tanjungpura
yohanes.gatot.sutapa.y@fkip.untan.ac.id
Sumarni
Universitas Tanjungpura
sumarni@fkip.untan.ac.id

Corresponding email: muslimdrmwn@gmail.com

Abstract

The primary purpose of this study is to find out information about the quality of item analysis on English teacher-made tests related to the difficulty level, item discrimination level, validity, and reliability of the final semester assessment for the twelfth-grade student of SMAN 8 Pontianak in the academic year 2020/2021. The researchers conducted this study using descriptive quantitative analysis with the data obtained from the English teacher-made test, consisting of 40 items with 257 test-takers of twelfth-grade students. The researchers analyzed the test items by using the combinations of Master TAP and SPSS Version 16 applications. Based on the difficulty level, the analysis results show that 22 items were categorized as easy, 14 items as medium, and four items as difficult level. Moreover, the item level of item discrimination shows that 9 items are categorized as poor, 18 items as enough, and 13 items as good index level. The analysis results also show that 23 items as valid questions, while the other 17 are invalid. On the other hand, the questions' reliability level showed satisfactory results, with the K-R20 result of 0.835. The results of the analysis showed that there were some questions that had already of good quality, but some problematic items were also found after being analyzed. Therefore, it is necessary for the teacher to conduct an evaluation of the problematic items so that the assessment process will be better in the future.

Keywords: English teacher-made test; Item analysis; Test-items

Introduction

In an educational setting, an assessment plays an important role. Shohamy (2017: 27-28) states that an evaluation is the very nature of language learning and its valuation and doing justice to providing value to the learning and students'

performance who are expanding their multi-/interlinguistic and multi-/intercultural ability. Moreover, Tosuncuoglu (2018: 163) defines assessment as a long-term procedure covering information and data concerning student development. It is an important activity for the teacher in the learning process to know how far students can master the materials to solve the problem given by the teacher in the form of a test and an assignment. Therefore, assessment is one of the vital procedures that are needed in the teaching and learning process.

The primary purpose of an assessment is to know the student's understanding of the material given and to determine whether or not they have achieved the learning objective. In conducting the assessment, the teacher needs to follow the guidelines set by the education unit. In Indonesia, the Ministry of Education and Culture (Kemendikbud) sets the criteria for learning. The 2013 Curriculum, revised in 2017, states the activities teachers can use to assess the students and what the teacher needs to prepare before evaluating the students.

The teacher needs to cover affective, psychometric, and cognitive aspects in preparing the assessment. Cognition is one of the most common aspects that must be assessed by evaluating students for tests or assignments. Rosli et al. (2016: 2) state that cognitive assessment is an instrument developed to evaluate the students' memory skills and how well they demonstrate other cognitive areas, including attention, decision-making, and language use. The cognitive test delivered to students comprises a daily assignment, mid-term test, and final semester assessment.

The final semester assessment is an activity in assessing students at the end of the semester, either held in the odd or even semester. This activity aims to measure the students' learning outcomes for one semester and the student's understanding of the teaching materials that the teacher delivers. The final semester assessment is a summative test. It is a part of an assessment to provide information about students' achievements and is usually taken at the end of the semester (Setiyana, 2016: 434). The final semester assessment is very crucial. Without any tests at the end of the semester, the teachers have difficulty measuring the students' mastery of the materials they teach. Therefore, the teacher must prepare a good quality test to assess the students at the final examination.

Test quality plays a significant role in increasing the student's learning outcomes. Samritin & Suryanto (2016: 94) state that a good quality test that is developed in accordance with the procedures of instrument development used to

obtain the information about students' capability progress. Accordingly, the teacher needs to analyze the test to determine the questions' quality. Item analysis is an activity to assess the quality of the items made and whether or not test items are acceptable, need revision, or are even thrown away. Moses (2017: 19) claims that item analysis is a statistical assessing procedure that the teacher needs to observe and improve items, estimate the characteristics of potential test forms, and conclude the quality of items and accumulated test forms. Moreover, Arora (2018: 6) states that to item analysis to an activity that examines students' responses to individual test items (multiple choice questions) to evaluate the quality of those items and the test as a whole. The items are analyzed after the students finish answering the tests. The test items can be analyzed manually or using a statistical program application.

In this study, the researchers takes the research at the twelfth grade of SMAN 8 Pontianak, located on Ampera street, Sungai Jawi, Pontianak. The researchers discussed the items used for the final semester assessment with the English teachers. Before implementing the test, the teacher already made guidelines used as references for the test items. From the observation, the results of the student's final semester assessment show that most of the students did not meet the standard criteria for English subjects.

The result shows that 121 of 257 students got lower than 75 for the final semester assessment, as the standard criteria for English subjects in SMAN 8 Pontianak. It means that almost half of the students did not meet the minimum standard of the subject. It indicates that the teacher needs to evaluate the test items. Unfortunately, the teacher of a twelfth-grade student in SMAN 8 Pontianak has not evaluated the test used to assess the students even though it was needed. It is considered important to evaluate to determine why the errors happened to prepare for better assessment in the future. Based on this problem, the researchers are interested in conducting this study on this school.

The previous studies on this issue included Humaerah (2016) and Toksöz and Ertunç (2017). The first researcher focused on analyzing a summative test's validity, reliability, and difficulty level. She examined the data manually. The researcher reported 60% valid items and 40% invalid items. Besides, she found out that the test items were reliable, and the level of difficulty was categorized; as complex (one item), too easy (one item), medium (four items), and easy (four items). The latter researchers focused on analyzing item facility, item discrimination, and distractor efficiency of the

multiple-choice test. They analyzed the data using IBM SPSS Version 20. They reported that most items were at the moderate level in terms of item facility, 28% of the test items were low at their discrimination value, and some distractors were significantly ineffective and should be revised.

This study focused on analyzing the level of difficulty, item discrimination, validity, and reliability of the test using the Master TAP application and SPSS version 16 to analyze the items. The test items to be analyzed in this research are applied to measure the students' command of English in the last year of their senior high school education. The present study examined whether or not the teacher-made test to measure the student's mastery meets the qualities to assess the students. Moreover, the present study is conducted to answer the following questions; (1) What is the level of difficulty of test items of the final test assessment English teacher-made for students of twelfth grade in SMA Negeri 8 Pontianak? (2) What is the level of item discrimination of test items of the final test assessment English teacher-made for students of twelfth grade in SMA Negeri 8 Pontianak?, (3) What is the validity of test items of the final semester assessment English teacher-made for students of twelfth grade in SMA Negeri 8 Pontianak?, and (4) What is the reliability of test items of the final semester assessment English teacher-made for students of twelfth grade in SMA Negeri 8 Pontianak?

Research Methodology

Research Design

The researchers used descriptive quantitative research as the design for this research. It refers to the study that focuses on describing a condition in the form of numbers that occurs factually and systematically. Descriptive research is a method used to reflect the existing phenomena as accurately as possible (Atmowardoyo, 2018: 198). At the same time, quantitative research is an approach to testing objective theories by examining the relationships among variables (Creswell & Creswell, 2018: 41). The former describes the data and characteristics of the observed phenomenon. The latter collects numeral data from a group of people, then generalizes the study results to explain a phenomenon.

Setting and Participant

The researchers conducted the study at SMAN 8 Pontianak, located on Ampera street, Pontianak. Moreover, the participant for this research is Year-12

students of SMAN 8 Pontianak consisting of 257 students as an informant for the researchers to get data used to analyze the test items. The data needed is the student's answer sheet on the final semester assessment.

Technique and Tool of Collecting Data

To elaborate and provide a solution to the research focus, the researchers used document review as the technique to collect data. The procedure of collecting the data in this research is the researchers took on the test indicators that the teacher used as the reference in making test items for the final semester assessment. After the examination, the researchers took the students' responses and the answer key of test items from the teacher. To make the analysis process more accessible, the researchers transferred the students' answer sheets from Google Forms into Microsoft Excel.

The tool for collecting data in this research is the English teacher-made test used for the final semester assessment. The researchers used secondary data from the teacher after getting permission to analyze the test items used to assess the students. The test consists of forty multiple-choice questions with five alternatives. The alternatives include one correct answer and four wrong answers. Before constructing the test, the teacher had finished making the guidelines used as a reference to establish the test. The test consists of materials for KD I until KD IV for Year 12.

The technique of Data Analysis

The data was analyzed quantitatively. Albers (2017: 5) states that quantitative data analysis is more than just discovering statistical significance; it is also linking the results of the statistical analysis with the relationship of the study and describing practical conclusions. The data was analyzed using the Master TAP application to determine the difficulty level, item discrimination level, and reliability. While for validity, the data was analyzed using SPSS version 16. The researchers analyzed the item following the procedure for using the application to find the best results of the analysis of the questions being tested.

After discovering the analysis results on all items, the researchers grouped the items by the suitable categories, according to the item classification index. In addition, researchers also took the average results of the analyzed items as a generalization of the item criteria. To present the results that are easy for the reader to understand, the

researchers presents the analysis results in a table, followed by a brief description of the item classification.

Findings and Discussion

Findings

Answering first research purpose: finding out the level of difficulty of test items of the final semester assessment English teacher-made for students of twelfth grade in SMA Negeri 8 Pontianak in the academic year 2020/2021. The following section provides more detailed information about the analysis results for the difficulty level.

Table 1: Analysis results for the level of difficulty

Index Classification	Number of Questions	Index Results	Notes
Easy (Questions with an index range of 0.70 < P ≤1.00)	1	0.84	The items in this category are generally the types of questions that have been mastered by students so that most students can answer the questions well. Therefore, it is necessary to evaluate the following questions so that the quality would be better to test students' understanding of the material that has been taught.
	2	0.87	
	6	0.81	
	7	0.73	
	9	0.95	
	10	0.88	
	12	0.72	
	15	0.81	
	17	0.84	
	18	0.72	
	20	0.73	
	21	0.91	
	23	0.75	
	24	0.85	
	25	0.91	
	29	0.81	
	30	0.97	
	33	0.72	
	34	0.82	
	37	0.82	
	38	0.79	
Medium (Questions with an index range of 0.30 < P ≤ 0.70)	4	0.60	The items in this category are generally good types of questions to test on students. Evaluation of these items only needs to be done if problems are found in other categories of item analysis.
	5	0.50	
	8	0.66	
	11	0.50	
	13	0.62	
	14	0.65	
	22	0.61	
	27	0.62	
	31	0.46	
	32	0.68	
	35	0.45	
	36	0.68	
	39	0.69	

	40	0.53	
Difficult (Questions with index range 0.00 $< P \leq 0.30$)	16	0.19	The items in this category include material that has not been mastered by the student. It is necessary to hold an evaluation.
	19	0.28	
	26	0.03	
	28	0.28	

In addition, the information about the second research purpose: finding out the level of item discrimination of test items, was obtained. The analysis results show that most items are categorized in enough index. It means that most students know how to answer the material being tested. The detailed information related to the analysis result of item discrimination level provides in the table below.

Table 2. Analysis results for the level of item discrimination

Index Classification	Number of Questions	Index Results	Notes
Poor (Questions with index range 0,00 – 0,19)	9	0.10	This is a category of questions that are less able to distinguish between students at low and high levels. Therefore, evaluating these questions is necessary to improve test quality.
	13	0.14	
	16	0.15	
	21	0.16	
	25	0.19	
	26	0.00	
	28	0.15	
	30	0.09	
	31	-0.04	
Enough (Questions with index range 0,20 – 0,39)	1	0.37	This is a category of questions that can distinguish students at low and high levels. The teachers can defend the following questions or evaluate if problems are found in another category of item analysis.
	2	0.38	
	3	0.22	
	5	0.33	
	8	0.28	
	10	0.22	
	12	0.37	
	17	0.29	
	19	0.28	
	20	0.24	
	22	0.39	
	23	0.36	
	24	0.27	
	27	0.39	
	29	0.25	
	35	0.30	
	37	0.38	
	40	0.33	
Good	4	0.46	This is a good question category for distinguishing students at low and high levels. Questions like this can be maintained, and evaluation is not required if there are no problems in other categories of item analysis.
	6	0.41	
	7	0.46	
	11	0.46	
	14	0.41	
	15	0.48	
	18	0.62	
	32	0.41	

(Questions with index range 0,40 – 0,69)	33	0.45	
	34	0.41	
	36	0.53	
	38	0.56	
	39	0.47	

Moreover, the third research purpose was to find out the validity of test items of the final semester English teacher-made assessment. The results show that 23 items for the final semester assessment are classified as valid questions, while 17 are classified as invalid questions. The specific information related to the analysis result on the validity level is described in the table below.

Table 3. Analysis results for the level of validity

Index Classification	Number of Questions	Results (R-Count)	R-Table	Notes
Valid (Questions with higher results than R-Table)	1	0.118	**0.113**	The questions in this category indicate that the items tested can perform their duties in measuring students' ability to teach materials. This means that the questions are good enough to be tested on students, and the evaluation of the questions can be carried out only if problems are found in other categories of item analysis.
	7	0.167		
	8	0.269		
	9	0.114		
	11	0.126		
	14	0.245		
	15	0.336		
	16	0.145		
	18	0.226		
	19	0.326		
	21	0.135		
	24	0.119		
	28	0.255		
	29	0.149		
	31	0.209		
	33	0.133		
	34	0.212		
	35	0.318		
	36	0.126		
	37	0.303		
	38	0.171		
	39	0.168		
	40	0.339		
Invalid (Questions with lower results than R-Table)	2	-0.042		The questions in this category indicate that the items tested are unable to perform their duties in assessing students' abilities. It is necessary to hold an evaluation or even an overhaul of the questions so that the questions become valid they can be tested on students.
	3	0.041		
	4	0.016		
	5	0.095		
	6	0.006		
	10	0.061		
	12	0.031		
	13	0.017		
	17	0.078		
	20	-0.049		
	22	0.049		
	23	0.090		

	25	0.083		
	26	0.109		
	27	0.073		
	30	-0.035		
	32	0.103		

Lastly, this section is to answer the fourth research purpose: finding out the reliability of test items of the final semester assessment English teacher-made. To find out the reliability level, the researchers analyzed the items using Kuder-Richardson (K-R20) formulas obtained from the test-items analysis results Master TAP application. The analysis results show that the test items' reliability was 0.835. The index level of reliability indicates that the test items being tested qualified as excellent reliability.

Discussion

A good test must be a test that meets the standards of assessment in the learning process, and the purpose is to measure what students have learned in the learning process following the learning objectives listed in the lesson plan. Therefore, a teacher was expected to make a test blueprint of questions as a reference in writing test items. Raymond & Grande (2019: 1) define that the test blueprint describes the parts that would be involved in a test, which is also followed by other important characteristics such as the emphasis given to each topic and the format of the test assessment. Based on the test blueprint as the guidelines for an assessment, the teacher carries out a final examination for the students to determine the extent to which students understand the material better.

Based on the results shown in the final assessment process, the researchers are interested in research to test the quality of the test items. By combining two types of applications generally used to test the quality of exam questions, the researcher conducted an item analysis test on the items tested. There are four primary focuses on item analysis in this study, including the level of difficulty, item discrimination, validity, and reliability.

The first part discusses the analysis's results on the difficulty level. Musa, Shaheen, Elmardi, and Ahmed (2018: 1478) state that item difficulty index is a assess of the proportion of the total examinees who fulfilled an test appropriately, mostly known as the p-value. The results show that the average difficulty level of the questions is 0.679, meaning the level of questions is categorized as moderate. Based on the analysis results, this shows that the questions tested have a good level of testing. Boopathiraj & Chellamani (2013: 190) state that the optimal level of problem difficulty is 0.50 to

distinguish the achievement between students with high and low achievement. Generally, items of medium difficulty are preferred over much easier or more complex.

However, several questions have potential problematic items. Based on analysis results, there are some questions that are indicated as too difficult questions. For example, out of 257 participants who participated in this exam, question number 26 could be answered by only seven students. It means that the questions were complicated and the students have not mastered the learning objectives. That is why arrangement is needed for this question. In addition, there were any other questions that are found as difficult level items, such as question number 16 that is correctly answered by only 50 students, question number 19 by 73 students, and question number 28 by 72 students.

On the other hand, the analysis results also show the questions with too easy level. For example in item number 9, 243 students could answer correctly. It indicates that the students can understand the learning objective well and the teacher need to reconstruct this item with the one. Moreover, there were also another easy level question based on the analysis results on Master TAP application, like item number 25 and 31 are answered correctly by 235 students, also item number 30 by 250 students.

According to Boopathiraj & Chellamani (2013: 190), an item with a high difficulty value indicates a simple test item and might be a concept less worthy of testing. While an item with a low difficulty value shows that the teacher should review the problematic items to evaluate possible confusing language or the content needs re-instructions. It indicates that the teacher needs to improve the quality of these questions to achieve an excellent standard for the test items by evaluating the problematic test items. This activity expects to help the teacher make questions with preferable quality to test on students in the future.

The second part discusses the analysis's results on the level of item discrimination. According to Ado (2013: 1657), item discrimination refers to the percentage distinction in correct answers between the lower and upper level students. The analysis results show that several questions have potential problems, especially questions with poor index. For example, question number 31 with a -0.04 index score shows that lower-level students can answer better than higher-level students. The analysis showed that 38 students of lower-level answered correctly on question number 31. At the same time, students with high-level could only answer correctly as many as 36. The researchers also found the same result in question 26 with

a 0.00 index score. The number of students at low and high levels had the same ability to answer the questions, where two students from each different level could answer the question correctly.

With these results, evaluating questions with a low-quality discrimination index is necessary. The teacher can improve in selecting more varied multiple-choice options and testing students' thinking skills. With this improvement, the teacher is expected to provide better quality questions in various aspects to examine students in the future.

Although the analysis results show that there are still questions with a low-quality discrimination index, the questions tested by the teacher have a pretty good discrimination index, so they are categorized as good questions and deserve testing on students. Kubiszyn & Borich (2013: 228) state that there is no specific answer to state that a test has a good discrimination index value. However, some experts agree that at least the results of the index value analysis show the number 0.30. In contrast, some experts agree that the item discrimination ability considers adequate as long as the analysis results show a positive value. These test items meet the appropriate index quality standards with a 0.319 score as the average discrimination index.

The researchers discusses the analysis results on the validity level of the test items in this part. Cheng & Fox (2017: 65) state the validity objective in the classroom assessment is to meaningfully and accurately define assessment information (e.g., grades, scores, teachers' oral and written comments, student's observation of errors, and students recognitions about their learning-those remarkable eureka moments when students become conscious of and/or acknowledge their learning). It indicates that validity shows the item analysis results and provides an overview of the item.

Based on the analysis results on SPSS version 16, 23 items are declared valid, and 17 items are invalid. The study also showed that 3 of the 17 invalid items have negative results, which means that the questions aimed at measuring students' abilities did not fulfill their function in measuring students' understanding of the material. The problematic items are item number 2, with a validity score of 0.042, item number 20, with a validity score of 0.049, and item number 30, with a validity score of 0.035.

Based on the results, teachers need to make significant revisions or eliminate the questions above because they did not fulfill the criteria of the indicators that need to achieve on the test blueprints. That follows Kubiszyn & Borich (2013: 327), which state that a test with good content validity must be suitable for instructional objectives.

It is purposeful that the teacher as a constructor can master the analysis of validity items to create test items that work on the ability to measure what it is supposed to measure.

The last part discusses the results of the analysis on the reliability of the test items. According to Sürücü, & Maslaci (2020: 2707), reliability refers to the constancy of the assessing tool used and its consistency over time. In brief, it is the capability to measure instruments to give identical results when adjusted at different periods. The results of the Kuder-Richardson analysis (K-R20 formula) in the Master TAP analysis show the reliability of the questions of 0.835 and are categorized as reliable test items. Mahajan (2017: 68) states that reliability needs more than 0.9 in high-stakes scenarios (like licensing exams), although values of 0.8 or 0.7 may be acceptable in moderate settings. It means that the questions met the test's ability to consistently measure the measured target even though it has been tested many times in different situations and groups of people.

Although the reliability shows satisfactory results, the teacher can still improve several parts to get higher reliability results. The analysis results on the Master TAP suggest that the test length should be 0.79 times longer, with a total of 32 items having the same quality as those in the current test to get K-R20 reliability of 0.80. Teachers can also get K-R20 reliability of 0.90 by testing 1.78 times longer with 71 items of the same quality as the current test.

Improvements in the parts of test items that need correction expect to help teachers become more familiar with creating questions with a good level of item analysis, which will maximize the measurement process for students in the future. It certainly has a good impact on student achievement in accepting and understanding the material that they are taught. On the other hand, improving the quality of measurement in the assessment system can also enhance the quality of the education process.

Conclusion and Suggestion

Conclusion

Based on the discussion of the previous part, the researchers can conclude that the exam questions of twelfth-grade students of SMAN 8 Pontianak at SMAN have moderate difficulty with a pretty good level of item discrimination. In addition, the test items tested also have a high level of reliability. On the other hand, the level of validity

of the questions also shows that the percentage of valid questions is more significant than the invalid ones.

Overall, the questions are categorized as a good level of testing to assess students' abilities. Still, it is necessary to make improvements and change items in several parts, especially for the problematic items, such as items with negative results, too easy or too difficult items, items with poor index in discrimination level, and invalid items. That needs to be taken seriously by the teacher to produce questions with a good level of item analysis in the future.

Suggestions

Related to the study's results, the researchers described several teacher suggestions: (1). The research findings indicate that the test items made by the teacher are not all in a suitable category of questions to be tested on students because there are several problematic items on them. Therefore, the teacher must evaluate the guidelines for writing test items and questions with a potentially problematic level (2). In addition, the researchers also found several writing errors in the final semester assessment test items, such as errors in the use of punctuation, grammar, or word writing. Therefore, the teacher must hold periodic checks on the items made to correct errors (3). After that, the teacher expects to be able to analyze and test the questions to obtain a valid and reliable test, and (4). The teacher expects to evaluate problematic exam questions. The teacher can revise or eliminate inappropriate questions, both from the level of item difficulty, item discrimination level, validity, and reliability. Meanwhile, the teacher can maintain the questions with good quality for another examination.

Furthermore, the researchers are also aware of limitations in this study, such as choosing only one grade level at one school or limited criteria for selected analysis items. Therefore, the researchers suggests the other researcher who has an interest in the same topic as follows; (a) Since this study only analyzes several items, the other researcher can add more item analysis criteria that the researcher did not include. For example, the future researcher can consider exploring distractor analysis to get more specific results on what teachers need to revise, especially arranging the multiple-choice question option, (2) This study focuses on analyzing the final semester exam test in only one school. Therefore, the other researcher can take another kind of test to analyze, for example, the National Standard School Examination (USBN), which has

a broader range of respondents. The results can also be generalized to students at the same grade level. Besides, the other researchers can also consider choosing a different grade level of students, (3) Moreover, the researcher used SPSS version 16 to analyze the test items. Further researchers can consider using the latest version of the application or other similar applications to analyze the quality of the tested items on students and (4) Furthermore. Future researchers may also consider conducting research with different research methods. That is intended so that future researchers can obtain more detailed research results.

References

Ado, A. B. (2015). Item Analysis using derived Science Achievement Test Data. International journal of science and research (IJSR), 4 (5): 1655-1662.

Albers, M. J. (2017). Quantitative data analysis—In the graduate curriculum. *Journal of Technical Writing and Communication, 47(2)*, 215–233. https://doi.org/10.1177/0047281617692067

Arora, S. (2018). Item analysis. International Journal of Nursing Science Practice and Research, 4(2), 6–7. Retrieved from https://nursing.journalspub.info/index.php?journal=IJNSPR&page=article&op=view&path%5B%5D=813&path%5B%5D=0 (Accessed 16 March 2021).

Atmowardoyo, H. (2018). Research methods in TEFL studies: Descriptive research, case study, error analysis, and r & d. *Journal of Language Teaching and Research, 9(1)*, 197-204. https://doi.org/10.17507/jltr.0901.25

Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *Journal of Social Science & Interdisciplinary Research, 2(2)*, 189-193. https://r.search.yahoo.com/_ylt=AwrOun3mXt1ilZ4iYzJXNyoA;_ylu=Y29sbwNncTEEcG9zAzIEdnRpZAMEc2VjA3Ny/RV=2/RE=1658703718/RO=10/RU=http%3a%2f%2fwww.indianresearchjournals.com%2fpdf%2fIJSSIR%2f2013%2fFebruary%2f15.pdf/RK=2/RS=8M1tdzP954 (Accessed 14 May 2022).

Cheng, L., & Fox, J. (2017). Assessment in the language classroom: Teachers supporting, student learning. London, United Kingdom: PALGRAVE.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches (Fifth edition)*. United States of America: SAGE.

Humaerah, M. (2017). *Item Analysis of English summative test for second grade student of MAN 1 Tanete Bulukumba (Unpublished Thesis)*. Makassar, South Sulawesi: UIN Alauddin Makassar.

Kubiszyn, T., & Borich, G. D. (2013). *Educational testing and measurement: Classroom application and practice*. New York, United States of America: Wiley.

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University. Economic Series, 17(4)*, 59–82. https://doi.org/10.26458/1746

Moses, T. (2017). A review of developments and applications in item analysis. In R. E. Bennett, & M. v. Davier, *Advancing Human Assessment* (pp. 19–46. https://doi.org/10.1007/978-3-319-58689-2_2). Cham, Germany: Springer International Publishing.

Musa, A., Shaheen, S., Elmardi, A., & Ahmed, A. (2018). Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartoum University. Khartoum Medical Journal, 11(2), 1477–1486.

https://www.researchgate.net/publication/328583573_Item_difficulty_item_discrimination_as_quality_indicators_of_physiology_MCQ_examinations_at_the_Faculty_of_Medicine_Khartoum_University.

Raymond, M. R., & Grande, J. P. (2019). A practical guide to test blueprinting. . Medical Teacher, 41(8), 854–861.

https://doi.org/10.1080/0142159X.2019.1595556

Rosli, R., Tan, M. P., Gray, W. K., Subramanian, P., & Chin, A.-V. (2016). Cognitive assessment tools in Asia: A systematic review. *International Psychogeriatrics, 28(2)*, 189–210.

https://doi.org/10.1017/S1041610215001635

Samritin, S., & Suryanto, S. (2016). Developing an assessment instrument of junior high school students' higher order thinking skills in mathematics. Research and Evaluation in Education, 2(1), 92-107.

https://doi.org/10.21831/reid.v2i1.8268

Setiyana, R. (2016). Analysis of summative tests for English. *English Education Journal (EEJ), 7(4)*, 433–447. Retrieved from

https://r.search.yahoo.com/_ylt=Awr48m8_et1irlYj2h1XNyoA;_ylu=Y29sbwNncTEEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1658710720/RO=10/RU=http%3a%2f%2fww

w.jurnal.unsyiah.ac.id%2fEEJ%2farticle%2fview%2f5525/RK=2/RS=E7WH8DcXfpYhDi6SABYiRFZnwmo- (Accessed 14 March 2021)

Shohamy, E. (2017). *Language testing and assessment*. New York, United States of America: Springer Science+Business Media.

Sürücü, L., & Maslaci, A. (2020). Validity and reliability in quantitative research. Business & Management Studies: An International Journal, 8(3), 2694-2726.

Toksöz, S., & Ertunç, A. (2017). Item analysis of a multiple-choice exam. Advances in Language and Literary Studies, 8(6), 141-146. https://doi.org/10.7575/aiac.alls.v.8n.6p.141

Tosuncuoglu, I. (2018). Importance of assessment in ELT. Journal of Education and Training Studies, 6(9), 163-167. https://doi.org/10.11114/jets.v6i9.3443