Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 9, August 2021: 356-370

The Characteristics of Indonesia Second-semester Final Test for Eighth-grade Students

Suwarto Suwarto[a]
[a] Prof. Dr., Veteran Bangun Nusantara University, Faculty of Teacher Training and Education, Sukoharjo, Indonesia.

Abstract

This study aims to: (1) describe the item difficulty, (2) describe the item discrimination, (3) describe an alternative, and (4) know the reliability of the test. The population is 126 answer sheets. The 60 samples are taken randomly. Data collection used is interviews and documentary. Data is analyzed by using the Iteman version 3.00 program. The study's result showed: (1) The item difficulty ranges from 0.150 to 1.000. (2) The item discrimination ranges from -9.000 to 1.000. (3) The answer key is well. The test has 38.33% effective distractor s and 61.67% ineffective distractors, and (4) The test's reliability is 0.531.

***Keywords**: Item difficulty, item discrimination, test reliability.*

Endonezya İkinci Yarıyıl Sekizinci Sınıf Öğrencileri İçin Final Sınavının Özellikleri

Öz

Bu amaç: (1) madde ayrımcılığını, (3) bir alternatifi kaynak ve (4) testin güvenilirliğini bilmek. Nüfus 126 cevap kağıdıdır. 60 numune alınır. Tarihi veri toplama, röportajlar ve belgeseldir. İteman sürüm 3.00 programlanabilir analiz edilir. Çalışmanın sonucu gösterdi: (1) Maddenin zorluğu 0.150 ile 1.000 arasında değişiyor. (2) Ürün ayrımı -9.000 ile 1.000 arasında değişmektedir. (3) Cevap anahtarı iyi. Testte% 38,33 etkili çeldiriciler ve% 61,67 etkisiz çeldiriciler vardır ve (4) Testin güvenilirliği 0,531'dir.

***Anahtar Kelimeler:** Madde zorluğu, madde ayırt etme, test güvenirliği*

Introduction

The final exam of the semester is one of the requirements that must be done by students to get a higher level. The Florida Standards Assessment (FSA) and End of Course Exams (EOCs), Florida's current version of mandated testing, are being administered each year (Furgione, Evans, Russell, & Jahani, 2018). Quantitative research methods like surveys, interviews, questionnaires, tests, tutor observation check-lists, students' and learners' performance records experiments were utilized (Seitenov, Aubakirova, Fominykh, & Belenko, 2020). (Jandaghi, 2010) states that the test is the most important part for teachers to be able to evaluate their students. Thus, the teacher can obtain information on how far students understand the subject matter and improve the teaching and learning process. (Shomami,

2014) states that the purpose of the test is to provide information on student progress to ensure the extent to which learning objectives have been achieved and to review the effectiveness of the teaching and learning process. According to (H. D. Brown, 2003); (Crocker & Algina, 2008), a test is a series of procedures to measure a person's ability, knowledge, or performance. Roszkowski & Spreat (2011) also argue that the test is a systematic procedure to collect students' information. Teachers can also use tests to motivate and help academic students (Jandaghi, 2010; Lai, 2011). By testing students, they are indirectly motivated to study hard. According to (H. D. Brown, 2003), there are several components of the test: (1). The test method must be explicit and structured to qualify as a test. (2). The test must measure students' abilities. (3). The test must measure student performance.

The test results can be used as feedback for teachers to improve and evaluate the teaching and learning process, while, students can represent all their performance in the teaching and learning process. A good test must consist of good test items that meet the requirements based on the characteristics of the test, and it must provide real information that contains the smallest possible error (Mulianah & Hidayat, 2013). According to Surapranata (2004) and Nugiyantoro, Gunawan, & Marzuki (2002), each student will obtain a score consisting of three parts, the observed test score, the actual score, and measurement error, therefore, measurement error should be minimized to improve measurement quality. End-of-semester tests often use multiple-choice tests and each item must be good. Tests that have the smallest possible error will be able to measure student achievement accurately (Sireci, Thissen, & Wainer, 1991). The measurement results must be reliable, thus, the characteristics of the test must be analyzed accurately (Ackerman, 1992). States that the characteristics of the test must have an adequate level of item difficulty, item discrimination, and the function of the distractor. Also, the reliability of the test is important because it will provide reliable measurement results (Kane, 1986). According to Nugiyantoro, Gunawan, & Marzuki (2002), if the reliability of the test is high, the test will be able to measure it by minimizing the error score as small as possible so that the test can measure student achievement accurately. High test reliability, good test items, the test measurement results will be accounted for (Kane, 1986). This is supported by Surapranata (2004). He states that quantitative analysis is an analysis of the internal characteristics of a test through empirical data. The analysis of the quantitative internal characteristics is item difficulty, item discrimination, and reliability. Also, there are an alternative option and the correct answer as an answer key, and the effectiveness of the distractor. Therefore, making a multiple-choice test must be considered: item difficulty, item discrimination, alternative option, and reliability. Surapranata (2004) states that one of the objectives in analyzing test items is to improve tests that meet the requirements: namely (1) it can be used because it is proven to be good items supported by numerical data that are analyzed statistically, (2) it can be revised for bad test items, and (3) it can be deleted because the test items do not function empirically. According to (Masruroh, 2014), the teacher who analyzes the test items will be able to find out which items are good or bad. Therefore, by analyzing tests, the teacher can determine which items can be used and which items should be revised or deleted.

In this study, the researcher interviewed one of the Indonesian language teachers at SMP IT MTA Karanganyar, Central Java, Indonesia. Based on the interview, the teacher himself made the final semester test in a Covid-19 pandemic situation. According to his admission, the test was also not tested. The teacher made a test grid then made test items and distributed them directly to students. Since the test was not tested, no empirical analysis was carried out. Therefore, the researcher was interested in

analyzing the eighth-grade student's second-semester final test. Based on the above background, the researcher formulated the following research questions: (1). What is the item difficulty of the second-semester final exam test items in Indonesian for eighth-grade students? (2). How is the item discrimination of the second semester final exam test items in Indonesian for eighth-grade students? (3). What are the alternative answers (distractor and answer key) in the second-semester final test of Indonesian for eighth-grade students? (4). How is the reliability of the second-semester final examination test in Indonesian for eighth-grade students?

Literature Review

Researchers found that teachers made tests without making grids first (Bijsterbosch, Béneker, Kuiper, & van der Schee, 2018; Coniam, 2014; Mohajan, 2017). They made tests based solely on their abilities. They didn't see the material and the syllabus. This was because they were not given direction from the test team leader in making a good test. There had been many studies conducted in the past to analyze the characteristics of the test, namely, item difficulty, item discrimination, alternative answers, and reliability.

First, the investigators analyzed the item difficulty or proportion correct (p). Researchers analyzed the quality of multiple-choice questions through item analysis (Singh, Kariwal, Gupta, & Shrotriya, 2014). They found 11 (55%) items that were in the medium category with a range of 30% - 70%, 9 (45%) items that were included in the easy category with a range of p> 70%, and no items that were included in the difficult category with a range of p <30 %. Chauhan calculated the item difficulty in a multiple-choice test on anatomy subjects (Chauhan, Ratrhod, Chauhan, & Rameshbhai, 2013). They found 35 out of 65 items included in the acceptable range (30-50% or 60-70%), 3 out of 65 items included in the difficult category (p <30%), 12 items out of 65 items included in the easy category (p> 70%), and 15 of the 65 items included in the ideal quality (50-60%). Most of the items are at an acceptable level of item difficulty. Suruchi & Rana analyzed the items on the Biology achievement test (Suruchi & Rana, 2012). They found 1 out of 120 items in the difficult category with p <0.20, 18 out of 120 items that were included in the good category with a range of 0.20 <p <0.50, 94 out of 120 items being the best category with the range of 0.50 <p <0.80, and 7 items out of 120 which are included in the very easy category with p> 0.80. Thus, they determined that one difficult item and seven easy items should be rejected for the final achievement test draft. Kolte found 4 difficult items with a range of p <30%, 26 acceptable items with a range of 30-70%, and 10 easy items with a range of p> 70% (Kolte, 2015). Sa'adah (2017) analyzed the quality of the test items for the mid-semester test of English. She found 18 items (72%) as ideal category with a range of about 0.62, 2 items (8%) as easy category with a range of p> 0.90, and 5 items (20%) as the difficult category with a range of p <0.20 (Sa'adah, 2017). Saputra compared the quality of tests for the second-semester test of English between SMP N 1 Semarang and SMP Kesatrian 2 Semarang (Saputra, 2015). He found 31 easy items, 15 medium items, and 4 difficult items from SMP N 1 Semarang, while at SMP Kesatrian 2 Semarang there were 36 easy items and 14 medium items. These researchers analyzed the item difficulty through certain formulas such as item difficulty (p), including that they analyzed manually, however, some researchers used a computer program to analyze it, for example, Mulianah & Hidayat used the Iteman program version 3.00 to analyze computer-based test items including item difficulty (Mulianah & Hidayat, 2013). Also, each researcher chooses a specific category theory to determine grain quality.

Second, the researchers analyzed item discrimination (D). It is usually calculated using a correlation index (Lababa, 2018). According to Crocker & Algina, there are four correlation indices used to calculate item discrimination: point biserial correlation coefficient, biserial correlation coefficient, phi coefficient, and tetrachoric correlation coefficient (Crocker & Algina, 2008). This is an item analysis that can be calculated manually or through computer software such as SPSS, Microsoft Excel, States program, and Iteman, version 3.00. For example, Chellamani & Boopathiraj have analyzed item discrimination using a separation method between the upper and lower groups whose scores are entered in Microsoft Excel (Boopathiraj & Chellamani, 2013). Zubairi & Kassim used SPSS and Bigsteps to analyze the characteristics of the items which were the item difficulty and the item discrimination (Ainol Madziah Zubairi, 2006). Another example, Raharja analyzed item discrimination with Anates V4. In his study, there was no very good category of item discrimination (Raharja, 2014). There were only 8 items in the good category, 13 items in the sufficient category, and 28 items in the bad category. Therefore, bad items should be removed, and sufficient items should be revised.

Third, the researchers analyzed the distractors. Distractors are wrong answer choices. The function of the distractor is to divert attention so that students are confused in choosing the correct answer. Distractors are said to be effective if selected at least 5% (0.050) of the respondents. Distractors are said to be ineffective if less than 5% of respondents are chosen. Distractors that are ineffective must be revised (Lababa, 2018; Mutaqi, 2007). Putri & Ujang analyzed the Iteman version 3.00 program, however, for reliability, item difficulty, and item discrimination, she analyzed manually (Putri, 2015). That shows that she didn't know that the Iteman program version 3.00 could analyze everything. This was also done by Rusmiana (Rusmiana, 2015).

A test is said to be reliable if the test is consistent over time to produce the same score. Reliability shows that the measurement results can be trusted. It means that a test must produce a reliable score. The use of measuring instruments repeatedly will give consistent results. This is supported by Harrys & Valette (2003) which states that reliability means the stability of the test scores. Crocker & Algina states that the consistency of test results is called reliability (Crocker & Algina, 2008). Suhr also states that the reliability of assessing is the accuracy and precision of the instrument (Suhr, 2003). Many studies conducted in the past found reliability using the Kuder-Richardson 20/21 formula (KR-20 / KR-21) (Bernasela, 2014; Haryudin, 2015; Pascual & North, 2016; Sugianto, 2017). Pascual & North illustrated that the English achievement tests for ESL students in the Northern Philippines are reliable (Pascual & North, 2016). However, there is a researcher, Hidayati who found moderate reliability in the mid-semester test of English for eighth-grade students of SMPN 33 Semarang (Hidayati, 2009).

The difference between these studies and this research is the data analysis technique used by the Iteman program version 3.00 to analyze and reveal the item difficulty, item discrimination, distractors and even answer keys. However, it can also be a similarity because Rusmiana (Rusmiana, 2015) also uses the Iteman, version 3.00 program to analyze the test characteristics. The difference between Rusmiana's research and this research is the object of research. Rusmiana's study analyzes the field of accounting for vocational education, while, the object of this research is the final test of the second semester of Indonesian for eighth-grade students.

Methodology

Research Design

In this study, the researcher revealed the characteristics of Indonesia's second-semester final test by analyzing the test item and reliability of the test, so the researcher used a descriptive quantitative approach. This research employed descriptive analysis because it was intended to reveal the characteristics of the test on the Indonesia second semester final test of the eighth-grade students of SMP IT MTA Karanganyar Central Java Indonesia. The researcher used quantitative research because numerical data analyzed statistically with the program Item and Test Program Analysis (Iteman) version 3.00 program.

Population and Sample

The population in this study were all 126 answer sheets of the second-semester Indonesia final test. This answer sheet was obtained from 4 classes, namely: Class 8A = 32 students, 8B = 32 students, 8C = 31 students, and 8D = 31 students. The sample of 60 sheets was taken randomly from the 126 sheets.

Data Collection

This study used two data collecting techniques, they were: interview and documents. The interview was conducted by the researcher to collect data. Firstly, the researcher asked permission to research the school to the school principal and administration. Secondly, the researcher asked one of the Indonesian teachers to get information about the curriculum of the school program and the data of whole eighth-grade students, then, ensure time for the researcher to take the data (Indonesia second final test paper, students' answer sheets). The documents were the Indonesia second final test papers, answer key, and students' answer sheets. From these answer sheets, they would be analyzed of each item test about the item difficulty, the item discrimination, the alternatives, and reliability. The test was also be analyzed to obtain a reliability index.

The technique of Data Analysis

The multiple-choice test and its answer of Indonesia's second final test at SMP IT MTA Karanganyar central Java Indonesia were analyzed to find out whether each item is easy, moderate, or difficult for the students to do, it is for item difficulty. For item discrimination, whether the quality of each item is bad, sufficient, good, and very good. Whether each item has good distractors or not and a good answer key. It was not only that but also to find out whether the multiple-choice test was reliabel. They would be analyzed by using Iteman version 3.00 program.

Item Difficulty

To find the item difficulty of each test item, the following formula:

$$p = \sum B/N \quad \text{.....(1)}$$

Where:

P = proportion of correct

$\sum$ B= the number of correct answers

N= the number of respondents. (Lababa, 2018).

The item difficulty can be classified into three that are easy, moderate, and difficult. According to Mutaqi (2007), the category of item difficulty is as follows:

Table 1.

The Category of the Item Difficulty

P = The item difficulty	Category
P > 0.700	Easy
0.300 ≤ p ≤ 0.700	Moderate
P < 0.300	Difficult

Based on the Iteman version 3.00 program, the item difficulty can be described through column Prop. Correct which could be seen from the output file of Iteman version 3.00 program. Prop. Correct is the proportion of students who answered correctly. The item difficulty index close to 0 or 1 showed the item is too easy or too difficult for students (Hayat, Pranata, and Suprananto, 1997).

Item Discrimination

Item discrimination is calculated with biserial correlation and point biserial correlation. Biserial correlation formula. To find out the item discrimination of each test item with biserial correlation formula. The formula that can be used to calculate the item discrimination index as follows:

$$r_{bis} = \frac{M_P - M_T}{s_T} \cdot \frac{P}{Y} \quad \text{.....(2)}$$

Where:

r_{bis} = biserial correlation coefficient

Mp = the criterion score mean of those who answered the item correctly

M_T = the criterion score means of all examiners

S_T = standard deviation

P = the proportion of examiners who answered the item correctly

Y = ordinate of the standard normal curve at the z-score associated with the *P*-value for this item. (Crocker, L., and Algina, 1986; Mutaqi, 2007).

Point biserial correlation formula. To find out the item discrimination of each test item point biserial correlation formula. The formula that can be used to calculate the item discrimination index as follows:

$$r_{pbi} = \frac{M_p - M_t}{S_T}\sqrt{\frac{p}{q}} \quad \text{.....(3)}$$

Where:

r_{pbi} = point biserial correlation coefficient

M_p = the mean criterion score for those who answer the item correctly

M_t = the mean criterion of total score

S_T = standard deviation of total score

p = *proportion of correct*

q = *proportion of false* (q = 1 – p) (Crocker, L., and Algina, 1986).

In Iteman version 3.00 program, biser correlation and point biser correlation can identify item discrimination (Hayat, Pranata, and Suprananto, 1997). For statistically, the researcher used point biserial correlation formula to calculate the item discrimination because many teachers used the formula (Rudyatmi & Rusllowati, 2017).

Also, Suwarto (2018) stated that the point-biserial correlation is a bivariate correlation technique. To use the technique, variable 1 is discreet data (dichotomous data), and variable 2 is continuous data (interval data). This technique is usually used to calculate item discrimination by correlating between item score and total score. The point biserial correlation coefficient (rpbi) is a statistical measurement used to estimate the degree of relationship between a dichotomous nominal scale and an interval scale (J. D. Brown, 2001).

The item discrimination can be classified into four that are bad, sufficient, good, and very good. The bad item is eliminated. The sufficient item should be revised, however, good and very good items are accepted and saved in the test bank (Mulianah & Hidayat, 2013).

Table 2.

The Category of the Item Discrimination

r_{bis} = Item Discrimination	Category
$r_{bis} \leq 0.200$	Bad
$0.200 < r_{bis} \leq 0.400$	Sufficient
$0.400 < r_{bis} \leq 0.700$	Good
$r_{biss} > 0.700$	Very Good

Alternatives Analysis

Alternatives analysis have two kind options, namely, answer key and distractors. An answer key is said good key if the biser and the point biser index of answer key are greater than biser and the point biser index of other options. Key needs to be checked if the biser and point biser keys are smaller than biser and the point biser index of other options. It can point out that the key is a problem. Based on the key column which is the star sign of the output of Iteman version 3.00, there are some indicators to point a key problem out (Hayat, Pranata, and Suprananto, 1997). Firstly, there is the question mark (?) of the column. Secondly, an item that has a key problem will appear imperative and declarative sentence that is "Check the key A was specified, D works better". It indicated that it needed to cross-check its answer key. Option A was the original answer key, but many students chose option D as the true answer. Thirdly, the index of biser and point biser column can analyze whether the answer key is good or not. The index of those two columns must have the highest index than other indexes of each column. Nevertheless, if one of the indexes in each column is not the highest index than other indexes, so it indicates that the answer key must have a problem (Hayat, Pranata, and Suprananto, 1997). Thus, if there is one of the indicators above, the answer key must be a problem, and it must be checked cross what is wrong with the answer key. It might be related to other options, its question, the right answer itself, and why many respondents tended to be interested in choosing the option. Distractor analysis. The distractor is said to be effective if it is selected a minimal 5% (0.050) of the respondents. Distractors are said to be ineffective if it is selected by less than 5% of respondents. The ineffective distractors should be revised. Lababa (2018) stated that distractors that do not fulfill the criteria should be replaced or revised with other distractors that may be more interesting and confusing for students to choose.

Reliability Tests

The researcher found reliability index with Alpha Crobanch (α) formula because Iteman version 3.00 program which could be seen on the last page of the output of Iteman version 3.00 program used alpha to point the reliability index out. Not only Iteman version 3.00 program used the formula, but also the Indonesia second semester final test is an instrument which has its answer is scala (dichotomous). The answer just has two answers that are a true answer (score 1) and a false answer (score 0). This formula can be used to calculate the scala dichotomous (Nugiyantoro, Gunawan, & Marzuki, 2002). As for the Alpha Cronbach (α) reliability coefficient formula:

$$\alpha = \frac{\mathrm{n}}{\mathrm{n}-1}\left(1 - \frac{\sum Vsubtests}{Vtest}\right) \quad \text{.....(4)}$$

Where:

$\alpha = Alpha - Cronbach$

k = the number of subtests (Cronbach, 1951).

The reliability index ranges from 0-1. A test is said reliable if the reliability index upper 0.700. The highest reliability coefficient of a test is close to index 1. It indicates that a test has perfect reliability (Roszkowski & Spreat, 2011; Rudyatmi & Rusllowati, 2017).

Table 3.

The Summary of the Characteristics Category of Multiple-Choice Test

Characteristics	Category	Criteria
Item Difficulty	Easy	P > 0.700
	Moderate	0.300 ≤ p ≤ 0.700
	Difficult	P < 0.300
Item Discrimination	Bad	Biser ≤ 0.200
	Sufficient	0.200 < Biser ≤ 0.400
	Good	0.410 < Biser ≤ 0.700
	Very Good	0.710 < Biser ≤ 1.000
Distractors	Ineffective	Elected < 5% teste
	Effective	Elected ≥ 5% teste
Answer Key	Recheck	Biser & point biser *key* < biser & point biser distractor
	Good	Biser & point biser *key* > biser & point biser distractor
Reliability	Unreliable	< 0.700
	Reliable	> 0.700

Findings

The lowest item difficulty (Prop. Correct) was 0.150 (item 2) and the highest item difficulty was 1.000 (item 1 and item 11). From these data, it can be concluded that the most difficult item is item 2, while, the easiest item was item 1 and item 11. The complete data are summarized in Table 4.

Table 4.

Category of Item Difficulty

Category	Items	Total
Easy (Prop. Correct > 0.700)	1, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 17, and 18	13
Moderate (0.300 ≤ Prop. Correct ≤ 0.700)	3, 15, 16, 19, and 20	5
Difficult (Prop. Correct < 0.300)	2 and 8	2
	Total	20

From Table 4 it can be seen that the item difficulty included in the easy category is 13 items. The item difficulty that was classified as easy was 13/20 x 100% = 65%. There were 5 items in the difficult category. The item difficulty that was classified as moderate was 5/20 x 100% = 25%. There were 2 items in the difficult category. The item difficulty that was classified as difficult is 2/20 x 100% = 10%. From the percentage of item difficulty for each category, it can be seen that most of the item difficulty was an easy category, while, the least item difficulty level is the difficult category with 10%.

The lowest of the item discrimination (Biser) was -9.000 (item 11) and the highest of the item discrimination were 1.000 (item 1, and item 13). Items that have negative discrimination must be dropped, namely item 11. The complete data are summarized in Table 5.

Table 5.

Item Discrimination Category

Category	Items	Total
Bad (Biser $\leq$ 0.200)	11 and 14	2
Sufficient (0.200 < Biser $\leq$ 0.400)	3, 5, 8, 9, 12, 16, and 18	7
Good (0.400 < Biser $\leq$ 0.700)	2, 6, 10, 15, 17, and 20	6
Very Good (0.700 < Biser < 1.000)	1, 4, 7, 13, and 19	5
	Total	20

From Table 5 it can be seen that there were 2 items in the bad category of item discrimination. The item discrimination that was classified as bad is 2/20 x 100% = 10%. The items classified as bad should not be used (Biser <0,200 must be dropped). 7 items were categorized as enough. The item discrimination that was classified as sufficient was 7/20 x 100% = 35%. The discrepancy of the items which was classified as sufficient still needs to be revised. 6 items are categorized as good. The item discrimination that was classified as good is 6/20 x 100% = 30%. 5 items were categorized as very good. The item discrimination that was classified as very good was 5/20 x 100% = 25%. From the percentage of item discrimination in each category, it can be seen that most of the item discrimination is sufficient.

Distractors are answers to multiple-choice questions which are wrong answers. Distractors should be the answers of students who have misconceptions, formula errors, or calculation errors. Thus, when students choose alternative answers, they will be confused. This happens to students who have low competence. Distractors can be effective distractors or ineffective distractors. Effective distractors will be selected by more than 5% of respondents (0.050 of respondents). It shows that in the Iteman version 3.00 output, Prop. Endorsing is ≥0.050. While, the distractor that is ineffective must be less than 0.050 (Prop. Endorsing <0.050). The output of the Iteman, version 3.00, a distractor that effectively functions and a distractor that doesn't work effectively on the second semester of the Indonesian test has been summarized as in Table 6. In Table 6, we can understand that there were 37 ineffective distractors. They need to be revised. There were 23 effective distractors. The percentage of ineffective distractors was 37/60 X 100% = 61.67%. The percentage of effective distractors was 23/60 X 100% = 38.33%.

Table 6. *List of Distractors*

Item	Ineffective Distractors	Effective Distractors	Answer Key
1	A, B, C	-	D

2	C, D	B	A
3	-	B, C, D	A
4	B	A, D	C
5	C, D	B	A
6	C, D	B	A
7	B, C	D	A
8	B	C, D	A
9	B, C, D	-	A
10	A, C, D	-	B
11	B, C, D	-	A
12	C, D	A	B
13	A, C, D	-	B
14	A, C, D	-	B
15	C	B, D	A
16	-	A, B, C	D
17	A	C, D	B
18	B, D	A	C
19	A	B, D	C
20	C, D	A	B
Total of distractor	37	23	
Total of item	18	2	

All answer keys had functioned properly because from the analysis results did not appear Check the key. The reliability of the Indonesia second semester final examination for eighth-grade students was 0.531. This reliability was classified as unreliable because Alpha was less than 0.700.

Discussion and Conclusion

Percentage of item difficulty easy: moderate: difficult was 65%: 25%: 10%. The results were almost the same as Pranania Safira's research, the quality of the items was based on the item difficulty aspect, namely, the analysis results showed that the item difficulty was the easy category (Safira, 2016). Likewise, the research results of Haryudin found 16 easy items (53.33%) of English summative tests (Haryudin, 2015). It was the category of item difficulty with the most questions in his study. Masruroh's study (2014) also found 70% of easy items were dominated in the analysis of summative English tests for second-grade students.

Item discrimination that can be used as a good test is item discrimination> 0.400. Meanwhile, from the results of the analysis, it turns out that there were 2 items of bad item discrimination, 7 items of sufficient item discrimination, 6 good item discrimination, and 5 very good item discrimination. Thus, there were only 11 items that have met the minimum requirements from the point of item discrimination's view. 9 items did not meet the requirements from the point of Item discrimination's view. Rudyatmi & Rusilowati (2017) state that bad items must be dropped, sufficient items must be revised, good items and the very good item can be stored in the question bank. Therefore, 2 bad items had to be dropped and 7 sufficient items should be revised. There were 11 out of 20 items that meet

the minimum requirements to be deposited in the question bank. The 11 items consist of 6 good items and 5 very good items. These 5 excellent points were the same as Rusmiana's research. He found 5 items included in the very good category. Then, the items that have not met the requirements were 18 out of 40 items. The 18 items were 9 bad items and 9 sufficient items (Rusmiana, 2015). Putri & Ujang also found 10 test items that had the bad distinction (Putri, 2015). Her findings were almost the same as the results of this study.

An effective distractor was a distractor chosen by the respondents at least 5% or 0.050 (Lababa, 2018; Mutaqi, 2007). The results of the distractor analysis showed that 23 distractors were effective (38.33%) and 37 distractors were ineffective (61.67%). 37 ineffective distractors should be revised. The ineffective distractors of this study were almost the same as Shomami's research, it was found that 34 (17%) of the distractors were effective and 166 (83%) were ineffective (Shomami, 2014). It means that he found more distractors that were ineffective than effective distractors.

The reliability of the test is not reliable with 0.531. It was still less than 0.700. It was the same as the findings of research conducted by Pranania Safira (2016), that the reliability coefficient of the final semester examination was less than 0.600 (Acun, Dem, & Nur, 2010). The reliability test should be at least 0.700 (Aslan & Aktaş, 2020; Sokip, 2019; Widoyoko, 2010; Yeşilçınar & Çakır, 2020).

The conclusions of this study are: (1). The item difficulty of the test ranges from 0.150 to 1,000. The level of hardness of the most difficult item was item 2 and the level of the easiest item was item 11. The ratio of the percentage of easy items: medium: difficult items was 65%: 25%: 10%. (2). The item discrimination of the test from -9.000 to 1.000. The lowest of item discrimination was item 11 and the highest of item discrimination was item 1 and item 13. 2 items have bad item discrimination. There were only 7 items that have item discrimination. 6 items have good item discrimination. 5 items have very good item discrimination. (3). Distractors were effective by 38.33%, while, distractors were ineffective by 61.67%. The answer key was all good, and (4). The reliability of the Indonesian second-semester test was 0.531.

Suggestions

Suggestions for this research: (1). The formulation of tests needs to be improved regarding the percentage of easy: medium: difficult, namely: 25%: 50%: 25%. (2). It is also necessary to improve the item discrimination that designs the test, namely items that have item discrimination above 0.400. (3). It needs to increase the reliability of the test above 0.700.

Statements of Ethics and Conflict of Interest

“I, as Corresponding Author, declare and undertake that in the study titled as “The Characteristics of Indonesia Second-semester Final Test for Eighth-grade Students”, scientific, ethical and citation rules were followed; Turkish Online Journal of Qualitative Inquiry Journal Editoral Board has no responsibility for all ethical violations to be encountered, that all responsibility belongs to the author and that this study has not been sent to any other academic publication platform for evaluation.”

References

1. Ackerman, T. A. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity From a Multidimensional Perspective. Journal of Educational Measurement, 29(1), 67–91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x
2. Acun, İ., Dem, M., & Nur, İ. R. (2010). The Relationship between Student Teachers' Citizenship Skills and Critical. Journal of Social Studies Education Research, 1(1), 107–123. https://doi.org/10.17499/jsser.38718
3. Ainol Madziah Zubairi, N. L. A. K. (2006). Classical and Rasch analyses of dichotomously scored reading comprehension test items. Malaysian Journal of ELT Research, 2, 1–20.
4. Aslan, M., & Aktaş, K. (2020). The relationship between learning strategies and achievement goal orientations of high school students. Elementary Education Online, 19(1), 400–414. https://doi.org/10.17051/ilkonline.2020.661869
5. Bernasela. (2014). An Analysis on English Summative Test Items. Tanjung Pura University.
6. Bijsterbosch, E., Béneker, T., Kuiper, W., & van der Schee, J. (2018). Characteristics of test items focusing on meaningful learning: A case study in pre-vocational geography education in the Netherlands. European Journal of Geography, 9(1), 62–79.
7. 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.
8. Brown, H. D. (2003). Language Assesment Principles and Classroom Practice. In United States of America: Pearson Education.
9. Brown, J. D. (2001). Statistics Corner: Questions and answers about language testing statistics: Point-biserial correlation coefficients biserial correlation coefficients. Shiken:The Japan Association for Language Teaching Testing & Evaluation Special Interest Group Newsletter.
10. Chauhan, P. R., Ratrhod, P., Chauhan, R., & Rameshbhai, G. (2013). Study of Difficulty Level and Discriminating Index of Stem Type Multiple Choice Questions of Anatomy in Rajkot. Biomirror, 4(6), 1–4. Retrieved from www.bmjournal.in
11. Coniam, D. (2014). Pursuing the qualities of a "Good" test. Frontiers of Education in China, 9(2), 238–249. https://doi.org/10.3868/s110-003-014-0018-x
12. Crocker, L., and Algina, J. (1986). Introduction to Classical and Modern Test Theory. In New York: CBS College Publishing. Retrieved from http://www.mich.gov/documents/mde/3_Classical_Test_Theory_293437_7.pdf
13. Crocker, L., & Algina, J. (2008). Chapter 7 - Procedures for estimating reliability. In Introduction to Classical and Modern Test Theory.
14. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555
15. Furgione, B., Evans, K., Russell, W. B., & Jahani, S. (2018). Divided we test: Proficiency rate disparity based on the race, gender, and socioeconomic status of students on the florida US history end-of-course assessment. Journal of Social Studies Education Research, 9(3), 62–96. https://doi.org/10.17499/jsser.27986
16. 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. https://doi.org/10.22460/p2m.v2i1p77-90.167
17. Hayat, B., Pranata, S. S., & Suprananto. (1997). Manual Item and Test Analysis (ITEMAN). Jakarta: Pusat Penelitian dan Pengembangan Sistem Pengujian, Balitbang Dikbud.
18. Harrys & Valette. (2003). Principles Language Testing I. New York: Mc Graw Hill.
19. Hidayati, A. D. (2009). the Analysis of Validity , Reliability , Discrimination Power and Level of Difficulty of First Mid-Term Test in the Case of the Eighth Grade Students of SMP 33 Semarang Faculty of Languages and Arts.
20. Jandaghi, G. (2010). Assessment of validity, reliability and difficulty indices for teacher-built physics exam questions in first year high school. Educational Research and Reviews, 5(11), 651–654. https://doi.org/10.4172/2151-6200.1000016
21. Kane, M. T. (1986). The Role of Reliability in Criterion-Referenced tests. Journal of Educational Measurement, 23(3), 221–224.
22. Kolte, V. (2015). Original article Item analysis of Multiple Choice Questions in Physiology examination . Indian Journal of Basic and Applied Medical Research, 4(4), 320–326.
23. Lababa, J. (2018). Analisis Butir Soal dengan Teori Tes Klasik: Sebuah Pengantar. Jurnal Pendidikan Islam Iqra', 5, 29–37. https://doi.org/10.30984/jpii.v2i2.538
24. Lai, E. R. (2011). Metacognition : A Literature Review Research Report. In Research Reports.

https://doi.org/10.2307/3069464
25. Masruroh, H. Z. (2014). An Item Anaalysis on English Summative Test for Second Grade Students of MAN Tulungagung 1 in Academic Year 2013/2014. A Script: State Islamic Institute Tulungagung.
26. Mohajan, H. K. (2017). Two Criteria for Good Measurements in Research: Validity and Reliability. Annals of Spiru Haret University, 17(3), 58–82. https://doi.org/10.26458/1746
27. Mulianah, S., & Hidayat, W. (2013). Pengembangan Tes Berbasis Komputer. Kuriositas, 2(6), 27–43.
28. Mutaqi. (2007). Analisis Butir Soal Terhadap Instrumen Evaluasi Kegiatan Diklat. Materi Workshop Direktur Diklat Di UDIKLAT PT PLN (PERSERO) Semarang, 1–10.
29. Nugiyantoro, B., Gunawan, Marzuki. (2002). Statistik Terapan Untuk Penelitian Ilmu-Ilmu Sosial. Yogyakarta: Gadjah Mada University Press.
30. Pascual, G. R., & North, C. (2016). Analysis of The English Achievement Test for ESL Learners in Northern Philippines. International Journal of Advanced Research in Management and Social Sciences, 5(12), 1–5. Retrieved from www.garph.co.uk
31. Putri, N. S. (2015). An Analysis of English Semester Test Items based on The Criteria of A Good Test for The First Semester of The First Year of SMK Negeri 1 Gedong Tataan in 2012/2013 Academic Year. A Script: Lampung University.
32. Raharja, N. S. (2014). Analisis Butir Soal Ujian Akhir Sekolah Produktif Pemasaran Kelas XII Pemasaran SMK Negeri 9 Semarang. Economic Education Analysis Journal, 3(3), 564–569.
33. Roszkowski, M. J., & Spreat, S. (2011). Issues to consider when evaluating "tests". In Financial planning and counseling scales, 13-31. Springer New York.
34. Rudyatmi, Ely., & Rusilowati, A. (2017). Evaluasi Pembelajaran. Semarang: Faculty of Mathematics and Science Unnes.
35. Rusmiana, F. D. (2015). The Test Item Analysis of 1st Semester Final Test of The Accounting Theory for Vocational Education: Case Study of SMK YPKK 1 Sleman Academic Year of 2014/2015. A Thesis: Yogyakarta State University.
36. Sa'adah, N. (2017). the Analysis of English Mid-Term Test Items Based on the Criteria of a Good Test At the First Semester of the Eighth Grade Students of Mts . Mathalibul Huda Mlonggo in the Academic Year of 2016 / 2017. Jurnal Edulingua, 4(1), 45–57.
37. Safira, P. (2016). Analisis Butir Soal Ulangan Akhir Semester Gasal Mata Pelajaran Bahasa Indonesia SMP Negeri 2 Magelang Tahun Pelajaran 2015/2016 (Vol. 9). https://doi.org/10.5151/cidi2017-060
38. Saputra, R. W. (2015). The Comparison Between The Second Mid-Term English Tests for The Seventh Gradersmade by The State and Private School Certified English Teachers (The case of test items analysis of SMP N 1 Semarang and SMP Kesatrian 2 Semarang in the academic year of 2013). Journal of English Language Teaching, 4(1), 1–5. https://doi.org/10.15294/elt.v4i1.7920
39. Seitenov, A. S., Aubakirova, R. Z., Fominykh, N. I., & Belenko, O. G. (2020). Technological dimension of pre-school teacher training at tertiary school: Fine arts concept-based case study. Journal of Social Studies Education Research, 11(2), 186–203.
40. Shomami, A. (2014). An Item Analysis of English Summative Test. A Script: Syarif Hidayatullah State Islamic University.
41. Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the Reliability of Testlet-Based Tests. Journal of Educational Measurement, 28(3), 237–247.
42. Sokip. (2019). Emotive behavior control to reduce intolerance and depression among secondary school students in tulungagung Indonesia. Journal of Social Studies Education Research, 10(4), 75–96.
43. Sugianto, A. (2017). Validity and Reliability of English Summative Test for Senior High School. Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature, 3(2), 22–38. Retrieved from http://ejournal.kopertais4.or.id/mataraman/index.php/efi/article/view/3191/2432
44. Suhr, D. (2003). Reliability, exploratory and confirmatory factor analysis for the scale of athletic priorities. Statistic and Data Analysis. Retrieved from http://www2.sas.com/proceedings/sugi28/274-28.pdf
45. Surapranata, S. (2004). Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Test. Bandung: PT Remaja Rosdakarya.
46. Suruchi, S., & Rana, S. S. (2012). Test Item Analysis and Relationship Between Difficulty Level and Discrimination Index of Test Items in an Achievement Test in Biology. Paripex - Indian Journal Of Research. https://doi.org/10.15373/22501991/june2014/18

47. Suwarto. (2018). Statistik Pendidikan. Yogyakarta: Pustaka Pelajar.
48. Widoyoko, E. P. (2010). Evaluasi Program Pembelajaran. In Yogyakarta: Pustaka Pelajar.
49. Yeşilçınar, S., & Çakır, A. (2020). Development and validation of the english teachers' attitudes towards recruitment system scale. Elementary Education Online, 19(3), 1548–1546. https://doi.org/10.17051/ilkonline.2020.733191.