LADU: Journal of Languages and Education
2022 VOL. 2, NO. 2, 67–76

p-ISSN 2774-2083
e-ISSN 2774-194X

OPEN ACCESS

Examining teacher-made English test in a language school

Taufiq Effendi[1], Ilza Mayuni[2]
[1]Universitas Gunadarma, Jawa Barat, Indonesia; [2] Universitas Negeri Jakarta, Jakarta, Indonesia

ABSTRACT

Background: Multiple-choice, teacher-made English tests have constantly been popular due to their immediate alignment to classroom instructions. However, ample studies have indicated the need for continuous evaluation of their quality to allow evidence-based feedbacks for sustained betterment of assessment practices.
Purpose: This study sought to examine the quality of a multiple-choice, teacher-made English formative informal assessment for four classes of high school students of an English course in Madura, Indonesia.
Design and methods: Data were collected from the test results of eighty students and put in an excel document. The data were then analysed with a computer application called Conquest to analyse the responses of each of the students on every item of the test. Based on this item response analysis, it turned out that the test could have achieved a higher credibility if necessary, moderations had been taken.
Results: The findings recommend that schools as well as teacher institutions need to provide necessary trainings to ensure in-service teachers and pre-service teachers possess adequate test development and test analysis expertise for continuous improvement of the learning, teaching and assessment practices.

Keywords: English classroom assessment; multiple-choice test; validity; reliability; item response analysis

Introduction

Multiple-choice, teacher-made English tests have constantly been popular in Indonesia. Indeed, it is the teachers who are expected to know the best about the process of their classroom instruction. They are also at the same time the most responsible parties to monitor the learning progress of their students.

Teaching and testing are inseparable for they complete each other. Teaching requires testing to see how well the teaching has been and testing requires teaching to make sure the test takers are well-prepared. As highlighted Effendi & Suyudi (2017), teachers normally handle large classes but are given limited time to perform non-teaching tasks. As a result, multiple-choice test construction has many times been the solution to this matter.

Multiple-choice test is both time and cost-efficient. Not requiring a rigorous rubric, this type of assessment instrument minimises the degree of unreliability as that of performance-based or subjective tests. To this end, countless anecdotal evidences reveal that subjective test is as well popular but to an extreme rarity, is without a well-constructed and accountable rubric which eventually results in a severe unreliability.

It is true, however, that a poor multiple-choice test is subject to speculation. It is speculative in a way that there is hardly a certainty to anyone if a correct response of a test

CONTACT Taufiq Effendi ✉ taufiq.effendi@gmail.com

taker was as a result of their subject mastery or expert guessing for the test provides the correct answer though along with its distractors. Studies have shown that multiple-choice tests are often problematic when it comes to the quality of its distractors in every single item.

The distractors are not sufficiently diverting the test takers so that the correct answer remains distinctive though independent from a mastery of the tested subject. It is also well-noted that good and plausible distractors are often unavailable (Hughes & Hughes, 2020). Furthermore, cheating is very likely facilitated (Hughes & Hughes, 2020). Due to its speculative nature, such test by no means causes a problem when interpreting the results of the test (Weir, 1998). It is therefore hard to come to a solid conclusion of the true mastery of the test takers.

In contrast, multiple-choice tests are claimed to have been categorised as objective tests. It was argued that subjective judgement of an examiner or more is disattached (Miller et al., 2013) for scoring merely involves matching every response with the provided answer key and then the correct responses are calculated based on the percentile rank as the standard norm. Still, what makes the correct answer is often highly subjective and therefore problematic in that it is subjectively determined by the teacher who develops the test (Weir, 1998). Furthermore, the distractors that are hoped to sufficiently divert the test taker are as well subjectively based on the teacher. Studies have discovered that typical problems of a multiple-choice tests are the inadequate plausibility of the distractors and problematic or controversial correct answer. This causes disagreement among the test developer, scorer and the test takers on which option serves as the best answer.

Studies have been conducted in Indonesian context to learn the credibility of English teacher-made tests. Hakim & Irhamsyah (2020) concluded that a test administered to several classes in a high school in Aceh Tenggara they examined showed satisfactory validity. However, their methodology was inadequate for content-related validity was the sole quality aspect they studied. Other qualities even more profound remained under-investigated. Quite similarly, Indrayani et al. (2020) claimed that an English summative teacher-made tests administered to all students in a public junior high school in Bali were regarded to demonstrate a relatively high quality. Whereas, the researchers did not look into the results of the tests as to learn the test takers' responses on every single item and option which would lead them to see the reliability, construct validity, criterion-related validity, item discriminability, distractors plausibility. The researchers, on the other hand, sought to examine the alignment of the tests with a list of desired norms which were apparently independent from the results of the administered tests. This resonates the conclusion of studies conducted by Santy et al. (2020) and Septi et al., (2020) who matched English teacher-made tests in schools in Bali with desired norms. They, as well, concluded that the test was very good despite the absence of confirmed responses of the test takers that would allow clarification to their assumed conclusion. Another criticism falls on the coder reliability which remained concealed. Thus, it was insufficient to claim that the tests in these studies were credible.

Different from previous studies, Jannah et al. (2021) examined the quality of English trialing test in a public school in Tangerang through an item response analysis. Their study used clear evidences of the responses of the test takers. These responses removed unconfirmed assumptions typically generated from sole test booklet analysis without the presence of the responses. This study uncovered that the test administered to more than 300 students for two consecutive academic years mostly contained moderately difficult items, yet half of the test inadequately discriminated between high and low-achieving students and therefore the test insufficiently revealed accurate representation of the students' true understanding. It was however uninformed if the tests were developed by a single individual

teacher or a committee of teachers which still indicates the need for necessary test development training. In contrast, a study by Lebagi et al. (2017) found out that English teacher-made test at a private primary school in Semarang was highly reliable and possessed high discriminability despite its controversially easy nature. Yet, the outcome of the interview confirmed that the students had intensive exposures to the language allowing them to score relatively high on the test. These two studies with confirmed results of teacher-made English tests foreground conflicting outcome where one study discovered credible test but the other discovered contrasting quality.

Another study with confirmed responses of the test takers was done in Sekolah Indonesia in Kuala Lumpur, Malaysia. It was conducted by Kholilah (2016) who discovered that the English teacher-made test was somewhat in need for improvement. The test had generally poor discriminability and insufficiently well validity. Its reliability, to a surprise, was under-investigated despite the promise of examining it in the research methodology section of the article. In terms of content validity, distractors plausibility and level of difficulty, the test was sufficiently good. The weak aspects, however, indicate that again test development expertise of the teacher is not yet sufficient. Similarly, a study conducted in an English Education program in Bali by Paramartha (2017) discovered that an English reading teacher-made test could hardly discriminate among students with different levels of mastery. Almost half of all the items mostly require removal and some revision. Still, half of the items contain implausible distractors. These studies highlight the urgency to carry out continuous evaluation of the quality of teacher-made multiple-choice tests.

Given the findings of previous studies, multiple-choice, teacher-made tests are commonly unescapable. Culturally, multiple-choice tests have been the most familiar type for Indonesian population in general in that they have been used as high stake examinations for more than half a century (Effendi & Suyudi, 2017). For teachers in particular, the tests are often the most attractive for they are seemingly easy to develop and to score. Fast scoring offers the most profound advantage to teachers who are confronted with time constraints (Wu & Adams, 2007). In Indonesia, for instance, where teachers generally teach large classes for at least 24 hours aweek , not to mention extra teaching hours many teachers do in other institutions in order to earn more to make a living, MC tests are incredibly helpful. For this reason, the expertise in developing a credible multiple-choice test and analysing its quality serve as a paramount essential that teachers need to possess.

A credible test will allow teachers to reveal accurate data of their students' ability. The speculative nature of the test can be significantly reduced and anticipated. Without the ability to reveal accurate information about the students' progress and real ability, there is no validity nor reliability which prove to be the two most fundamental capital of a test. A credible test generates credible results and eventually leads to accurate and appropriate treatment and better classroom instructions. This is extremely essential considering that the general quality of English instructions in the country has not been satisfactory yet (Dardjowidjojo, 2000; Effendi, 2021; Sukyadi & Mardiani, 2011; Yulia, 2014).

In this present study, the quality aspects that were examined include reliability, construct validity, discrimination power, and distractors efficiency. Reliability serves as one of a key aspect of a test quality. It is commonly defined as a test result consistency (Bachman & Palmer, 1996; Brown, 2004; Hughes, 2012; Weir, 2005). That means, a good test is a test the result of which is relatively unchanging regardless its second or more attempts. The cut-off score of a good reliability reaches the value of 0.70. In other words, if a test has a lower reliability value, it is likely unreliable for the result of the testis very likely to change upon more attempts. This means, if the same test is done by the same group in two or more

different times, the result of the test would likely vary. But a reliable test would show otherwise.

The quality of reliability can be established in two conditions. The first one can be in a form of constructing similar items consecutively. This is referred to as item homogeneity (Hughes, 2012). These similar items are meant to measure the same single sub-construct. A capable test taker would certainly have correct answers on these consecutively given similar items. Thus, the result would not vary which means consistent. On the other hand, an incapable test taker would certainly speculate and end up with inconsistent responses for they may have a correct answer on one item but an incorrect answer on the next similar item. Therefore, a capable test taker would consistently score the same high result while their counterparts would consistently score lower. However, an unreliable test would not allow this consistency for it fails to discriminate the capable from their counterparts.

Another condition is by preparing two sets of the same test. This test constitutes set 1 and set 2. Each of these sets are identical or similar in terms of the order of the sub-constructs being tested, the number of the items, as well as the distractors. Set 1 will go from item number 1 to 50. Set 2 will go from 51 to 100. Item number 1 and 51 are identical or twins and so the rest of the twin items in two sets. This is called a parallel form or split-halves technique (Hughes, 2012). A capable test taker would have similar high score on both sets. Whereas, a less capable test taker would have relatively similar lower score on both sets. This technique is more time and cost-efficient and practical. This works in most cases where test-retest approach is unlikely to happen. This approach is to collect the results of the same test in different times administered to the same group.

It is worth noting that item homogeneity and split-halves are apparently synonymous with concurrent validation or validity. Concurrent validity is an attempt to verify the test takers' responses of two or more items measuring the same sub-construct (Messick, 1989; Alderson, 1991a). This equivalence drove Weir (2005) to consider reliability as a scoring validity. This is because the score or the result of a test significantly influences the interpretation of the result assumed to represent the achievement of a test taker. If the score is unreliable, the interpretation of the result will certainly be misleading. This is why scoring validity or reliability is by no means an integrated part of the overall validity.

As the second aspect examined, construct validity, together with content and criterion validity, is a pivotal quality. It is to ensure that the test measures the intended construct. Construct itself refers to a concept, knowledge, ability a test taker is expected to possess. Another understanding of construct validity is how sufficient is a test result understood and used (Bachman & Palmer, 1996; Effendi & Suyudi, 2017; Miller, Linn & Gronlund, 2009). These two understandings, in other words, highlight that construct validity is not only about the appropriateness of the test items but also about the appropriateness of the conclusion and functionality of a test score. If a test is construct valid, people will correctly know test takers' ability or mastery based on their score. In contrast, if a test is construct invalid, people will misunderstand test takers' real ability or mastery based on the score. This is why, by implication, if a test is construct invalid, the score will mislead the use of the test score.

What is also important is discrimination ability of a test. Discrimination ability refers to how a test can discriminate between high-achieving and low-achieving test takers and even among different levels of ability. It is very critical that people need to know which students or test takers possess high ability and which ones possess low ability. This can be seen from the discrimination power value (Hughes, 2012). Discrimination power value, as generated by Conquest, is shown for every individual item. This signifies that a good test employs items that can differentiate among test takers in terms of their achievement.

Another equally essential aspect from a multiple-choice test is its distractors efficiency. A Distractor is an option that is equally selectable as that of the correct answer of every individual item. Thus, a single test item has to be accompanied by one correct answer and three or more sufficiently distractible options. A good multiple-choice test must therefore employ attractable distractors that have relatively equal opportunity to be selected by the test takers. This, therefore, indicates the necessity not to develop distractors that appear very distinctly. Otherwise, it is unlikely that a test taker would choose them. It is true, however, coming up with plausible distractors are not easy for they are not always available (Hughes, 2012; Weir, 1990).

The above four qualities are essential. There is no hierarchical importance and neither greater weight on either one of them. All of them make up a solid body of a good test. Each of the aspects are influential for one another. Lacking one of them cause a test to suffer from a poor test that leads to a problematic validity and reliability.

This research aims to know how the reliability, construct validity, discriminability, and distractor efficiency of a multiple-choice, teacher-made English test developed for high school student's formative assessment in the targeted field of the research.

Methods

This study carried out an item response analysis. Such analysis is based on an Item Response Theory (IRT), a subclass of Psychometric, which is aimed to discover the relationship between test takers' answers to individual test items and the position of the test takers' performance on relevant continuum (Reckase, 2009). Brown (2004) highlights that examining item responses can help teachers as the test developers to discover the feebleness of the test they made and with this evidence, they are able to make necessary amendments. Having a high quality and credible test is critical to be able to generate accurate information about the true achievement of the students as the test takers. Analysing every single answer that the test takers chose allows us to clarify our assumption, see the quality of each of the distractors, the quality of every single test item, and the general quality of the test. Without analysing the test takers' responses, one would remain in their unconfirmed assumptions about the quality of the test. This information will allow teachers to have appropriate decisions about their future strategies.

Data collection and analysis

The study investigated the quality of an English test in an English course in Madura. The test was in a multiple-choice format and was developed by an English teacher. The test was administered as an informal formative assessment to four classes of senior high school students. The students' responses were stored in a Microsoft Excel document. Afterward, the data were analysed with an application called Conquest which is the product of the Australian Council for Educational Research. This application generates rigorous information about the credibility of a test administered to a limited group or even to a massive group.

Findings & Discussion

Reliability

Based on Conquest calculation, the outcome showed that the reliability value scored 0.48. This means that the reliability of the teacher-made test under-investigation was very slightly lower than what is expected from a teacher-made test. Frisbie (1988) points out that for

teacher-constructed test, the expected value is 0.50 while for the commercial ones, it is 0.90. The test being examined is lacking 0.02 value to achieve the acceptable reliability level. With little moderation, this test should be able to achieve the cut-off reliability score. This means that the test under-examination is relatively almost sufficiently meeting the standard reliability for a teacher-made test.

At present, the reliability of the teacher-made test understudy, as discovered, scored lower than that of other studies. It is below the reliability value of a test examined by Lebagi et al. (2017). The test in their study reached 0.92 which proves to demonstrate a high level of reliability. This level is even equal to a commercially-developed multiple-choice test. In Paramartha's study (2017), interestingly, the reported reliability score was the one generated after removing items that were problematic. The reliability score then passed the cut-off score. This was the result of the removal of 24 problematic items out of fourty items in total. With some moderations, the reliability of the test in this study can be enhanced significantly. As shown in Paramartha's study, problematic items can simply be removed or revised (Hughes, 2012). Another strategy is to add more items (Weir, 1990: p. 54-55). This attempt to increase the test reliability is only made possible if the teacher has the expertise in evaluating a test quality.

Construct validity

Conquest, the item analysis application, indicates the degree of construct validity with a term called "weighted MNSQ". Yet, Wu and Adams (2007). Report that weighted MNSQ refers to both construct validity and the suitability of every item to the item response model. The closer the score to 1.00, the higher the degree of the construct validity. The following table provides further illustration.

Table 1. Construct validity

Item	1	2	3	4	5
MNSQ	0.96 (0.87, 1.13)	1.04 (0.69, 1.31)	1.04 (0.88, 1.12)	1.03 (0.87, 1.13)	0.92 (0.59, 1.41)
Item	6	7	8	9	10
MNSQ	0.94 (0.70, 1.30)	0.97 (0.56, 1.44)	0.95 (0.85, 1.15)	1.03 (0.67, 1.33)	1.01 (0.86, 1.14)
Item	11	12	13	14	15
MNSQ	0.92 (0.24, 1.76)	1.04 (0.44, 1.56)	0.96 (0.44, 1.56)	1.03 (0.80, 1.20)	1.10 (0.87, 1.13)

It is worth pointing out that the ideal fit value and MNSQ is 1. This fit value shows the ideal discriminating power according to the prediction of the model (Wu & Adams, 2007: p. 66). MNSQ value of 1 indicates the desired construct validity. Yet, it is suggested that items with MNSQ value that is exceedingly higher than 1 have to be moderated.

As shown in the table, the first five items were in close approximation to the desired value of construct validity. The first item scored 0.96. The second and the third items scored slightly higher which is 1.04. The fourth item scored slightly lower which is 1.03. And the fifth item turned out to have the lowest score in the group which is 0.92.

Slightly different, the second five items were mostly below the value of 1. Item number 6, 7, and 8 had the value of 0.94, 0.97, and 0.95 respectively. Whereas, the last two items in the group scored slightly above the value of 1 which are 1.03 and 1.01 respectively.

In contrast, the last group of the items were mostly slightly above the construct validity value of 1. Items number 12, 14 and 15 had the values of 1.04, 1.03 and 1.10 respectively. Yet, items number 11 and 13 were lower than 1 which are 0. 9to and 0.96 respectively. Out

of all items, the last one seems to deserve a revision due to its score that most substantially surpassed the desired value.

The outcome of the analysis discovered that the fit value of all items falls within close approximation to the desired value. Many of the values are very close to the expectation while others are higher than the expectation but only slightly. These numbers inform that all items possess the desired construct validity for the measure the same construct.

The finding of this degree of construct validity of a teacher-made English test neither confirms nor contrasts the findings of other studies. Other studies in the field in Indonesian context, to the author's knowledge, did not seek to examine the construct validity of the test they investigated (Hakam & Irhamsyah, 2020; Indrayani et al., 2020; Jannah et al., 2021; Kholilah, 2016; Lebagi et al., 2017; Paramartha, 2017; Santy et al., 2020; Septi et al., 2020). It is with Conquest application that allows the generation of the degree of construct validity of the test. It is well-noted that validity is the alignment between what it aims to test with what is actually tested. This highlight the profound importance of examining the construct validity in every pursuit of test quality investigation. This way, the finding to some extent fills in some gap in the body of the research

Discrimination ability of individual item

This section discusses the discrimination power of the test. The value of the discrimination power of each individual item is provided in each column in the following table.

Table 2. Discrimination power

I	Disc.	Disc. ref.		Performance Reference					
	Index	Correct answer	Pt-bis	Distractor 1	Pt-bis	Distractor 2	Pt-bis	Distractor 3	Pt-bis
1	0.49	42 (52.50)	0.49	A: 8 (10.00)	-0.24	B: 3 (3.75)	0.05	C: 27 (33.75)	-0.38
2	0.26	63 (78.75)	0.26	A: 9 (11.25)	-0.19	C: 4 (5.00)	-0.03	D: 4 (5.00)	-0.19
3	0.29	39 (48.75)	0.29	B: 13 (16.25)	-0.05	C: 13 (16.25)	-0.19	D: 15 (18.75)	-0.13
4	0.33	44 (55.00)	0.33	A: 32 (40.00)	-0.28	B: 3 (3.75)	-0.10	D: 1 (1.25)	-0.09
5	0.51	12 (15.00)	0.51	A: 38 (47.50)	-0.22	B: 18 (22.50)	0.05	C: 12 (15.00)	-0.27
6	0.45	17 (21.25)	0.45	A: 26 (32.50)	-0.11	B: 18 (22.50)	-0.04	C: 19 (23.75)	-0.27
7	0.38	11 (13.75)	0.38	B: 5 (6.25)	-0.07	C: 42 (52.50)	-0.16	D: 22 (27.50)	-0.07
8	0.50	30 (37.50)	0.50	A: 26 (32.50)	-0.22	B: 11 (13.75)	-0.08	D: 13 (16.25)	-0.30
9	0.23	64 (80.00)	0.23	A: 5 (6.25)	-0.00	C: 2 (2.50)	-0.20	D: 9 (11.25)	-0.19
10	0.43	33 (41.25)	0.43	B: 11 (13.75)	-0.25	C: 21 (26.25)	-0.31	D: 15 (18.75)	0.03
11	0.52	5 (6.25)	0.52	A: 28 (35.00)	-0.09	C: 15 (18.75)	0.06	D: 32 (40.00)	-0.21
12	0.09	8 (10.00)	0.09	A: 19 (23.75)	0.18	C: 35 (43.75)	0.01	D: 18 (22.50)	-0.25
13	0.37	8 (10.00)	0.37	A: 43 (53.75)	-0.08	B: 6 (7.50)	-0.08	D: 22 (27.50)	-0.10
14	0.35	55 (68.75)	0.35	B: 10 (12.50)	-0.45	C: 14 (17.50)	-0.05	D: 1 (1.25)	0.06
15	0.14	34 (42.50)	0.14	A: 7 (8.75)	-0.38	B: 25 (31.25)	0.17	D: 14 (17.50)	-0.11

Notes: I = Item; Disc. = Discrimination; disc. ref. = discrimination reference; Pt. Bis= Point-biserial

The table presented above displays some important information. First, it tells how well an individual item can discriminate among the students based on their level of achievement. Second, it shows how many students selected the correct answer and how much percentage it is in the respective population. Third, it foregrounds how well every single distractor attracted the students. Number of selectors and its percentage are indicated.

In this analysis, there are two standards each signifying acceptable and high discrimination power. The lowest score to be considered acceptable is 0.20 while the cut-off score of a high discrimination power is 0.40 (Wu & Adams, 2007: p. 64). In other words, scores below 0.20 are insufficiently discriminating the test takers according to their ability.

As illustrated in the table above, six items were found to have possessed a high discrimination power. These items are number 1, 5, 6, 8, 10, and 11 which scored 0.49, 0.51,

0.45, 0.50, 0.43, and 0.52 respectively. The highest level of discrimination power in this group is 0.52 belonging to item number 11. Whereas, the lowest in this group is on item number 10 scoring 0.43.

The majority of the items, seven out of fifteen, were found to demonstrate acceptable level of discrimination power. This is shown in items number 2, 3, 4, 7, 9, 13 and 14 each scoring 0.26, 0.29, 0.33, 0.38, 0.23, 0.37 and 0.35 respectively. The two highest scores in this group are 0.37 and 0.38 which are slightly below the threshold to be in the group of high discrimination power.

On the contrary, the outcome figured out that at least two items deserve a modification. Items number 12 with 0.09 and number 15 with 0.14 demonstrate unacceptable discrimination power. This means, the two items cannot help people to know which students are high performing and which ones are not based on their responses on these items.

Given the two problematic items, this study show some disagreement and correlation with some other studies. Jannah et al. (2021 discovered that the teacher-made test they examined had roughly half of all the items that could not discriminate the test takers on their ability. Likewise, a study conducted by Kholilah (2017) revealed that even more than half of all the test items had poor discrimination ability which could hardly inform who had a better and who had less ability. Paramartha (2017), on the same token, found many items on the examined had low discrimination power. On the contrary, this present study resonates very well with the study by Lebagi et al. (2017) who uncovered that most items of the teacher-made English test under investigation proved to be strong in discriminating the test takers on their ability. This, to some extent, highlights some understanding that teacher-made English tests may or may not strongly discriminate students on their ability scale.

The degree of the plausibility of the distractors

In general, it turned out that every single distractor formulated in each item was indeed distractible. Each of them, however, received different degree of attraction. Some distractors were chosen by a few students while other distractors were opted by more students. Still, some other were selected by even a greater number of students. Out of 45 distractors, there are at least a few of them that require moderation. They are the options b and d item number 4, option C item number 9, and option d item number 14. These distractors only attracted insignificant number of test takers which are between only one and two students.

As shown in table 2, it can be seen that distractor efficiency is not very much correlated with the degree of discrimination power. This is indicated by items number 12 and 15 which scored below the acceptable discrimination power. Yet, all the distractors were relatively attractive to some students. On the other hand, the less attractive distractors were in the items that significantly surpassed the threshold grade of acceptable discrimination power and even almost reached the cut-off score for high discrimination power.

This finding does not confirm the study of Jannah et al. (2021) who figured out that the distractors of the test they examined were less effective. They were not attractive enough to divert the students from the correct answer. Different from Jannah et al. (2021), Lebagi et al. (2017) indeed discovered that roughly 70% of all the distractors were attractive. Apart from these findings, Paramartha (2017) who examined teacher-made English reading test in higher education context discovered that half of the distractors were problematic and therefore could not divert the test takers.

Conclusion

At this stage, the study managed to uncover important information about the quality of the teacher-made English test in a language school in Madura, Indonesia. An item response

analysis with the aide of the application revealed that the reliability of the test was nearly sufficient. It was only 0.02 lower than the desired cut-off score of the reliability for a teacher-made test. Another outcome proved that most of the items possessed the expected construct validity. Yet, one item was in need for modification to achieve a more ideal score. The next outcome portrayed a positive picture of the discrimination power of the items. Many items had high discrimination power, most had acceptable power but two could not sufficiently discriminate the students based on their ability.

The final quest generated another positive information about the efficiency of the distractors. Although most distractors were quiet and very attractive to the test takers, a few of them were somehow quite insignificant. All these findings suggest that the test is mostly positive but somehow necessary moderations need to be made in order to enhance the overall quality of the test. This way of analysis allows a comprehensive investigation of which parts are already good and which parts are still problematic which require some revision. Based on this evidence, then, one would be able to come up with appropriate decisions in regards to the subsequent learning, teaching and assessment practices for a more maximum learning outcome. By implication, this study recommends that teacher institutions and educational institutions devote a special attention on the urgency of test development and test analysis expertise.

References

Alderson, J.C. (1991a). Dis-sporting Life. Response to Alistair Pollit's paper. In Alderson and North (eds.), pp. 60-7.

Bachman, L. F., & Palmer, A. S. (1996). *Designing Language Test.* Oxford: Oxford University Press.

Brown, H. D. (2004). Language Assessment: Principles and Classroom Practices. New York: Pearson Education Inc.

Dardjowidjojo, S. (2000). English Teaching in Indonesia. *EA Journal*, *18*(1), 22-30

Effendi, T. (2021). Past learners'voices on EFL classroom management in Depok, Indonesia. Asian EFL Journal, 28(1.2), 220-240.

Effendi, T., & Suyudi, I. (2017). The Impacts of English National Examination in Indonesia. In Ninth International Conference on Applied Linguistics (CONAPLIN 9) (pp. 236-239). Atlantis Press.

Effendi, T., & Suyudi, I. (2017). Investigating the Quality of a Popular Classroom Assessment Instrument in Indonesia. In Ninth International Conference on Applied Linguistics (CONAPLIN 9) (pp. 67-70). Atlantis Press.

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. Educational Measurement: Issues and Practice, 7(1), 25-35.

Fulcher, G., & Davidson, F. (2007). *Language, Testing, and Assessment.* New York: Routledge.

Fulcher, G. (2010). *Practical Language Testing.* London: Hodder Education.

Hakim, L., & Irhamsyah, I. (2020). The analysis of the teacher-made test for senior high school at State Senior High School 1 Kutacane, Aceh Tenggara. JURNAL ILMIAH DIDAKTIKA: Media Ilmiah Pendidikan dan Pengajaran, 21(1), 10-20.

Hughes, A. (2012). Testing for Language Teachers (2nd Ed.). Cambridge: Cambridge University Press.

Indrayani, M. S. D., Marhaeini, A. A. I. N., Paramartha, A. A. G. Y., & Wahyuni, L. G. E. (2020). The Analysis of the Teacher-Made Multiple-Choice Tests Quality for English Subject. Journal of Education Research and Evaluation, 4(3), 272-278.

Jannah, R., Hidayat, D. N., Husna, N., & Khasbani, I. (2021). An item analysis on multiple-choice questions: a case of a junior high school English try-out test in Indonesia. Leksika: Jurnal Bahasa, Sastra dan Pengajarannya, 15(1), 9-17.

Kholilah, N. (2016). The quality of english language testing implemented in kbri school, sekolah indonesia kuala lumpur, Malaysia. IJET (Indonesian Journal of English Teaching), 5(1), 149-172.

Lebagi, D., Sumardi, S., & Sudjoko, S. (2017). The Quality of Teacher-made test in EFL Classroom at the Elementary School and Its Washback in the Learning. Journal of English Education, 2(2), 97-104.

Messick, S. (1989). Validity. In R. Linn (ed.), *Educational Measurement.* New York: Macmillan, pp. 13-103.

Miller, M. D., Linn, L. R., & Gronlund, N. E. (2013). *Measurement and assessment in teaching 10th Ed.* New Jersey: Pearson Education.

Paramartha, A. G. Y. (2017). The Analysis Of Multiple-Choice Test Quality For Reading III Class In English Education Department, Universitas Pendidikan Ganesha Bali, Indonesia. Journal of Education Research and Evaluation, 1(1), 46-56.

Santy, N. P. L., Dewi, N. L. P. E. S., & Paramartha, A. A. G. Y. (2020). THE QUALITY OF TEACHER-MADE MULTIPLE-CHOICE TEST USED AS SUMMATIVE ASSESSMENT FOR ENGLISH SUBJECT. Prasi: Jurnal Bahasa, Seni, dan Pengajarannya, 15(02), 57-70.

Septi, N. K. D. C., Paramartha, A. G. Y., & Wahyuni, L. G. E. (2020). AN ANALYSIS OF THE QUALITY OF TEACHER-MADE MULTIPLE-CHOICE TEST USED AS SUMMATIVE ASSESSMENT FOR ENGLISH SUBJECT. Jurnal Imiah Pendidikan dan Pembelajaran, 4(2), 356-368.

Sukyadi, D., & Mardiani, R. (2011). The washback effect of the English national examination (ENE) on English teachers' classroom teaching and students' learning. *K@ ta*, *13*(1), 96-111.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment. 8*, 350-353.

Wu, M. & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Educational Measurement Solutions: Melbourne.

Weir, C. (1998). Communicative Language Testing. New York: Prentice Hall.

Weir, C. (2005). *Language Testing and Validation: An Evidence-based Approach.* New York: Palgrave Macmillan.

Yulia, Y. (2014). An evaluation of English Language Teaching programs in Indonesian Junior High Schools in Yogyakarta province. Unpublished PhD thesis. Melbourne: RMIT University.