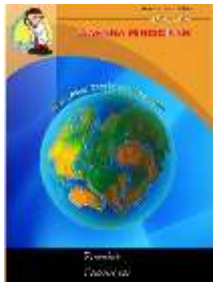
**Jurnal Ilmiah Wahana Pendidikan**

https://jurnal.peneliti.net/index.php/JIWP

**Vol. 7, No.8, Desember 2021**



# Analysis of Teacher-Made Tests Used in Summative Evaluation at SMP Negeri 1 Tompaso

**Nihta V.F. Liando*[1], Eunike Serhalawan[2], Ceisy Wuntu[3]**

[1,2,3]Universitas Negeri Manado, Tondano City, Indonesia
*Email: nihtaliando@unima.ac.id

**Info Artikel**
Sejarah Artikel:
Diterima: 22 Oktober 2021
Direvisi: 25 November 2021
Dipublikasikan: Desember 2021
e-ISSN: 2089-5364
p-ISSN: 2622-8327
DOI: 10.5281/zenodo.5775342

***Abstract:***
*The purpose of the present study is to describe the quality of English tests for use in summative evaluation in sixth semester in 2019-2020 and in fifth semester 2020-2021, The study was descriptive-and evaluative in nature in that it tried to determine the advantages (strengths) and disadvantages (weaknesses) of the summative tests used. in terms of their validity, reliability and item analysis. The data were the fifth and sixth semester students' responses to the two test in semester examinations. The data obtained were statistically analyzed using Point-Biserial for validity, KR-20 for reliability and item analysis for level of difficulty and discrimination power. Result of the statistical and item analysis indicated that (1) concerning the validity and reliability of the fifth semester 2020/2021 summative test, the data analysis shows that all the items are valid with validity coefficient ranges between 0.53 and 0.98, whereas reliability coefficient is 0.82, meaning that the test is highly reliable. It can be concluded that the test is valid and highly reliable, (2) concerning validity and reliability of the sixth semester summative test in 2019/2020, the analysis indicates that all the items are valid with validity coefficient ranges between 0.53 and 0.85, whereas reliability coefficient is 0.88, meaning that the test is highly reliable. It can be concluded that the test is good in its validity and highly reliable, (3) Item analysis item facility of the fifth semester 2020/2021 summative test shows that item facility index of the 50 items in the test ranges between 0.55 and 0.67, meaning that the test is good in terms of its item facility index; and of 40 multiple-choice items in the sixth semester 22019/2020 summative test, 27 items are considered very good items, and 13 items are reasonably good in their discrimination power. It can be concluded that the test is good in terms of its item facility and discrimination power, and (4) Item analysis of the sixth semester 2019/2020 summative test shows that in terms of item facility, all the 50 items are recommended for use. In terms of item discrimination, the analysis shows that 30 are very good or acceptable, and, therefore, are recommended for use, 19 are reasonably good, and only 1 is marginal and subject to improvement. It can be concluded that the test is good in its item facility and discrimination power. In general, the two summative tests, one used in the fifth semester 2020 and the other one in the sixth semester 2019 are good in terms of the characteristics of a good objective type test in multiple-choice format.*

***Keywords:*** *validity, reliability, item analysis, difficulty level, item discrimination*

## INTRODUCTION

In every walk of life the process of evaluation takes place in one or the other form. Take for example, if we want to buy a new cloth, consciously or not, we make a kind of evaluation of what we want to buy. In cooking, a chef usually takes the sample of what he/she is cooking in order to check whether it is as delicious as expected. If the evaluation process is eliminated from human life then perhaps the aim of life may be lost. It is only through evaluation that one can discriminate between good and bad. Therefore, it is said that the whole cycle of social development revolves around the evaluation process.

In education, evaluation is a must. Education is considered as an investment in human beings in terms of development of human resources, skills, motivation, knowledge and the like. Evaluation is described as the way of knowing students' achievement held at the end of teaching learning process Lalogiroth and Tatipang (2020, p. 2). Evaluation helps to build an educational program, assess its achievement and improve upon its effectiveness. It serves as an in-built monitor within the program to review the progress in learning from time to time. It also provides valuable feedback on the design and the implementation of the program. Thus, evaluation plays an important role in education.

In teaching and learning process, evaluation plays an enormous role. It helps teachers and learners to make decision to improve teaching and learning. It is a continuous process and a periodic exercise. It helps in forming the values of judgment, educational status, or achievement of students. In learning, it contributes to formulation of objectives, designing of learning experiences and assessment of learner performances because as stated in Lumentut and Lengkoan (2020,p. 20) learning itself is a system. Besides this, it is very useful to bring improvement in teaching and curriculum. Evaluation is concerned with assessing the effectiveness of teaching, teaching strategies, methods and techniques Pikirang, Liando, & Wuntu, (2021, p. 70). It provides feedback to the teachers about their teaching and the students about their learning. Stiggings (2004 cited in Saefurrohman and Balinas (2016, p. 83) puts it, 'Teachers should use assessment not only to actively and continuously measure a learner's progress but also to acquire useful data to inform their own instructional practice.' Bachman & Palmer (1996, p. 8) explained that language tests provide valuable information on various aspects of a language teaching-learning process which may be used to evaluate the teaching-learning program itself.

One of the most common instruments used in evaluation in English language teaching is test. Tests of various formats are used at schools to assess cognitive domain of students and skills. Since tests used in university admission, career promotion, semester and end-year exams are high stake in nature. Results of tests used in end-year exam, for example, are used to determine whether students advance to the next grade level or not. For this reason, test developers should make sure that the tests are properly developed. In this way, we are not in danger of doing something that is quite wrong – promoting people or holding people back on the basis of a wrong score. Tests used should measure what they are intended to measure (Field, 2005 in Taherdoost, 2016, p. 29).

Teachers' ability to develop high-quality test is highly important, because with the good quality test teachers are able to measure precisely the success of the learning process they have done. However, the quality of the teacher-made tests is sometimes in a big question whether or not they are well-done according to the underlying concepts of good tests and measure the instructional objectives as of instruction (He Lianzhen & LvZhouyang, 2013). Norris (2015) points out that the students have always complained that they are fed up with a test that is ambiguous,

unclear, and irrelevant. Some comment from students such as "I do not know what the teacher looks for and I studied the major details of the course but was only examined on the trivia and the footnotes. The problem is such a kind of test results in incorrect scores, and this, in turn, may result in serious consequences."

Although, as revealed by Lebagi *et al* (2017), the teacher-made test can be classified in good test, and the test brings both positive and negative washback in students' motivation in learning, other previous findings indicated that teacher-made tests still had a number of weaknesses. Saefurrohman and Balinas (2016) found that Indonesian junior high school English teachers used assessment for learning as the main purpose of assessment. Indonesian junior high school English teachers used items from published textbooks as their primary sources for constructing assessment items. Similarly, it was also found that teachers did not developed their tests based on test specification. They tended to copy-paste commercial tests or from exercise books, and did not develop their tests based on the principles of good evaluation. Teachers including English teachers were incompetence in developing their own test (see Survey by Direktorat Pembinaan Sekolah Menengah, Kementerian Pendidikan Nasional, 2010, p.105).

Problem related to ability to develop good quality of teacher-made tests is a common phenomenon. Badara (2016) found that the tests made by Indonesia language teachers are the teacher-made test invalid and unreliable. Mulyani (2020) also found that the test on financial accounting subject was not valid and reliable. Setiabudi et al (2019) arrived at the conclusion that the test still need some revision and improvement in order to be a good, valid, and reliable test. Rohmah (2018) found that content validity and reliability were considered good. However, some items were found to be either too easy or too difficult in terms of its index of difficulty. The test also had 50% poor discrimination index and 40% ineffective distractors

In the same vein, Yohana (2009) revealed that the test is not valid and need some revisions. The test makers should pay attention to the writing of multiple choice items and the characteristics of a good test. Primadani and Sulistyo (2014) revealed that the quality of the test is low in terms of the difficulty level, item discrimination, item validity, and the distracters even though the test reliability is acceptable and the items are 100% valid in covering the materials presented in the curriculum. To sum up, English teachers still find it difficult to develop high quality tests for use in semester exams.

## REVIEW OF LITERATURE

### What is test?

When assessing students reading comprehension, English teachers may use various kinds of methods or tools. One of the tools that commonly used to assess students' progress in a given subject, for English example, various kinds of tests can be used depending on the purpose of the test itself. A test is an instrument or procedure designed to elicit performance from learners with the purpose of measuring their attainment of specified criteria (Brown, 2001). Tests may be constructed primarily as devices to reinforce learning and to motivate the student, or primarily as a means of assessing the students' performance in the language (Heaton, 1975). Thus, a test can be used for both instructional and testing purposes.

Language tests can be classified in terms of purpose. When seen in terms of purpose, tests including language tests are classified into four categories, namely proficiency, placement, achievement and diagnostic tests. A proficiency test is usually used to assess leaner's general knowledge about a subject, English for example. TOEFL test is an example. Placement test is a test that is used to determine a student's level of ability in one or more subjects in order to place the

student with others of the same approximate ability. In the Standards for test construction (APA, 1999) achievement is viewed basically as the competence a person have in an area of content. Achievement test is a test that is used to assess students' mastery of instructional materials that they have already learned, while d*iagnosis* consists of identifying the nature of an illness or other problem through the examination of relevant symptoms. In language testing, a diagnostic test helps identify a student's learning problems so teachers can provide instruction to remedy those problems (Popham. 2009, p. 91).

In English language teaching context, language tests are used to assess the four major skills: listening, speaking, reading and writing. In addition, the tests can also be used to assess students' knowledge of language components, namely phonology, vocabulary, and grammar (Heaton, 1975).

**Characteristics of A good Language Test**

A test's usefulness, according to Bachman and Palmer (1996), can be determined by considering the measurements qualities of the test such as reliability, validity, practicality, discrimination and authenticity. These qualities can easily describe a good language test's usefulness. The test usefulness is the most important quality or cornerstone of testing. They state that test usefulness provides a kind of metric by which we can evaluate not only the tests that we develop and use, but also all aspects of test development and use.

A good language test should be valid. The term validity refers to "the extent to which the test measures what it says it measures" (Alderson & Hughes, 1981:135). For classroom teachers, content validity means that the test assesses the course content and the outcomes using formats familiar to the students. *Content validity* is defined as "the degree to which items in an instrument reflect the content universe to which the instrument will be generalized" (Straub, Boudreau et al. 2004). *Construct validity* refers to the 'fit' between the underlying theories and the methodology of the language learning and the type of assessment. For example, a communicative language learning approach must be matched by communicative language testing. Face validity means that the test looks as though it measures what it is supposed to measures. *Face validity* is the degree to which a measure appears to be related to a specific construct, in the judgment of non-experts such as test takers and representatives of the legal system (Taherdoorst, 2016, p. 29). Put it another way, face validity refers to one's subjective assessments of the presentation and relevance of the measuring instrument as to whether the items in the instrument appear to be relevant, reasonable, unambiguous and clear (Oluwatayo, 2012). Validity is an important factor for both students and administrators.

In addition to validity, a good test should be reliable. This means that the results of a test should be consistent (remain stable, should not produce different results when it is used in different days). A reliable test will yield similar results with similar group of students took the same test on two occasions, and their results are roughly the same — then the test will be called a reliable test. If the results are very different. Then the test is not reliable. A test is also reliable in the following cases:

a) If two comparative groups of students (students of similar abilities) score similar marks even if the test is given to them on two different days (provided that the students have not compared notes and prepared specially for it). If on the other hand, the results are so different, that in one group, the students score above average marks and the students in the other group fare badly, then the test is unreliable.
b) A test is reliable if students are marked by different teachers, and this does not produce high different marks.

c) Finally, a test is reliable if it has been properly administered. A 'perfect' test administration is one of that allows all examinees to perform at their best level under identical conditions. Conditions outside the test itself (e.g., the seating arrangement, bad acoustics, etc.) must not stop a student from performing at his / her best level. Thus a reliability has three aspects to it: reliability of the test itself, the reliability of the way in which it has been marked, and the reliability of the way in which it has been administered (Rajhy, 2014).

There are three aspects of reliability, namely: equivalence, stability and internal consistency (homogeneity). The first aspect, equivalence, refers to the amount of agreement between two or more instruments that are administered at nearly the same point in time. Equivalence is measured through a parallel forms procedure in which one administers alternative forms of the same measure to either the same group or different group of respondents. This administration of the various forms occurs at the same time or following some time delay.

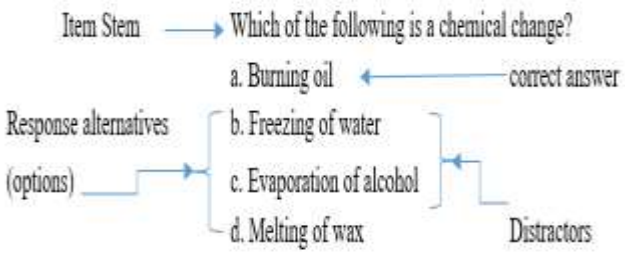
The second aspect of reliability, stability, is said to occur when the same or similar scores are obtained with repeated testing with the same group of respondents. In other words, the scores are consistent from one time to the next. Stability is assessed through a test-retest procedure that involves administering the same measurement instrument to the same individuals under the same conditions after some period of time. Test-rest reliability is estimated with correlations between the scores at Time 1 and those at Time 2. Two assumptions underlie the use of the test-retest procedure. The first required assumption is that the characteristic that is measured does not change over the time period. The second assumption is that the time period is long enough that the respondents' memories of taking the test at Time 1 does not influence their scores at the second and subsequent test administrations.

The third and last aspect of reliability is internal consistency (or homogeneity). Internal consistency concerns the extent to which items on the test or instrument are measuring the same thing. If, for example, you are developing a test to measure organizational commitment you should determine the reliability of each item. If the individual items are highly correlated with each other you can be highly confident in the reliability of the entire scale (Rajhy, 2014, p. 2–3).

**Multiple Choice Test**

In assessing students' performances, two types of assessment can be used: formative and summative. The Formative assessment provides immediate evidence of student learning, purpose to improve quality of student learning and promote modifications in curricular design, and the way we teach. The students receive individual feedback about their strengths and weaknesses. The Summative assessment occurs most frequently in the final of a course, semester or module. Essentially it is mostly used to make a final decision about the student performance.

In Both formative and summative assessment, multiple-choice test is most commonly used. The multiple choice item consists of the stem, which identifies the question or problem and the response alternatives or choices (or options). Usually, students are asked to select the one alternative that best completes a statement or answers a question. For example:




Multiple choice items are considered to be among the most versatile of all item types. They can be used to test factual recall as well as levels of understanding and ability to apply learning. As an example, the

multiple choice item below is testing not only information recall but also the ability to use judgment in analyzing and evaluating.

A Good multiple-choice item should not only be valid and reliable, but also be at the level of students' ability/knowledge, can differentiate those who learn and those who do not, and has distractors which can function effectively.

**Item Analysis**

Item analysis is the act of analyzing student responses to individual exam questions with the intention of evaluating exam quality. It is an important tool to uphold test effectiveness and fairness. Usually it consists of item difficulty, item discrimination, and effectiveness of distractor analysis. Here in this proposal, only item facility and item discrimination are briefly described.

**Item Facility Analysis**

Item facility (IF), also called item difficulty (ID) is a statistic used to examine the percentage of students who answer a given item correctly. It is calculated by adding up all correct answers in a given item, and divided by the total number of students who took the test. Item facility calculation follows this procedure: (1) count the number of correct answers in each item, (2) count the number of students who took the test, and (3) fill in the formula, and do the calculation.

An item facility value can range from 0.00 to 1.00. Thus, .45 indicates that 45 percent of the total students taking the test answer the item correctly. Normally, an IF range between .30 and .70 are usually considered acceptable (Brown, 2005:75).

**2.5.2 Item Discrimination**

Item discrimination (ID) is a statistic indicating the degree to which an item separates the students who performed well from those who did poorly on the test as a whole. The analysis can be done following these steps: (1) line up the students; names, their individual item responses, and the total scores in descending order, from the highest to the lowest based on the total scores, (2) proportionally divide the listing into three groups: upper, mid, and lower group, (3) count the number of correct answers of each item in the upper and/or lower groups, (4) divide the number of correct answers in each group with the number of students in the group, and (5) the ID is obtained by subtracting the IF of the lower group from the IF of the upper group. The guideline for making decisions is based on ID suggested by Ebel (1979:267) as depicted in Table 1.

Table 1. Criteria for Item discrimination

| ID Range | Meaning |
|---|---|
| .40 and up | Very good items |
| .30 to .39 | Reasonably good, but possibly subject to improvement |
| .20 to .29 | Marginal items, usually needing and being subject to improvement |
| Below .19 | Poor items, to be rejected or improved by revision |

## RESEARCH METHODOLOGY

**Research Design**

In terms of purpose, the proposed study is a descriptive-quantitative research in that it research involves collecting data in order to test hypotheses or to answer questions concerning the current status of the subject of the study (Gay et al, 2012). More specifically, the study tried to find out whether or not the items in the multiple-choice tests used in semester examination of the fifth and sixth semesters in the a academic year 2019-2020 and 2020/2021 at SMP Negeri 1 Tompaso were good in terms of their validity, reliability, item difficulty, and discrimination power.

In terms of purpose, this study can be classified as an evaluative study. Moore (1983:338) puts it, "The outcome of an evaluative study is to determine the advantages (*strengths*) and disadvantages (*weaknesses*) of a particular educational program, process, or product in terms of

specific objectives and criteria." Briefly put it, the purpose of this study is to provide information concerning the quality of English tests used in fifth and sixth semester examinations at SMP Negeri 1 Tompaso in the academic year 2019/2020 and 2021.

**Source of Data**

The data in this study were students' responses to the multiple-choice tests administered in the sixth and fifth semester examinations at SMP Negeri 1 Tompaso in the academic year 2019/2020 and 2020/2021. Their responses were in the form of test scores on the basis of which validity, reliability, item difficulty and discrimination power were analyzed. The responses to be analyzed were taken from two classes, one in the academic year 2019/2020 and another one in 2020/2021 academic year.

**Data Analysis**

Since the data in this study were numerical in nature, the data analysis was statistically done. The first was the analysis of test validity. The test to be analyzed was an objective type test in multiple-choice format. Therefore, item validity was done using Biserial correlation. This statistical technique is usually used to analyzed objective type test which is scored in which the correct answer

The multiple-choice tests to be analyzed in the present study were criterion-referenced in that they were administered in semester exam to obtain data concerning students' mastery of the instructional materials that had already been taught. However, since the data (scores) were also used as the basis to determine position each of the test-takers in comparison to other test-takers in the group, the item analysis was done using statistics commonly used for norm-referenced tests which are based on the *concept of normal distribution* (see Brown, 2005).

Empirical validity was analyzed statistically based on the type of data collected. Discreet data (data obtained using objective type test) were analyzed using Point Biserial correlation. The interpretation of the correlation coefficient is based on the general rule: the higher the coefficient, the higher the validity of the test item.

Table 2. Criterion for determining test validity

| Criterion | Meaning |
|---|---|
| **≤ 0.3** | Not valid |
| **≥ 0.3** | Valid |

Test reliability of each test was analyzed using split-half reliability technique in which half of the test items (odd number) were correlated with another half (even number). "If you have dichotomous items (e.g., right-wrong answers) as you would with multiple choice exams, the KR-20 formula is the best accepted statistic" (Dr. Korb). Therefore, reliability of the tests will be analyzed using KR-20. Coefficient of reliability (KR-20) was interpreted based on the following criteria put forward by (Sutrisno Hadi, 1979:310 & Weiresma and Jurs, 1990).

In addition to validity and reliability, two kinds of item analysis were also carried out: *item facility* and *item discrimination.* Item facility or item difficulty or *IF* for short was analyzed using this statistic:

$$\text{Item Facility} = \frac{\text{No. of correct response}}{\text{Total no. of students}}$$

Item discrimination index, or ID I for short, was calculated using this statistic:

$$\text{ID} = \frac{\text{No. correct in higher 27\% group - No. correct in lower 27\% group}}{\text{No. in higher group or lower group}}$$

Some experts put forwards their comments on ideal item facility. Oller (1979, p. 247) shows items falling somewhere between .15 and .85 of IF values are usually preferred. Brown (1996,

p. 70) states that those fall in a range between .30 and .70 if IF values are said to be applicable. According to International Assessment Resources (2011), item facility or item difficulty ranges between 0.20 - .09. In this study, it is interpreted based on the criteria put forward by Musa et al (2019, p. 1478)

Regarding item discrimination, Oller (1979:252) indicates that usually a value of .25 or .35 is set as a lower limit on acceptable IDs. In this study, the interpretation of item discrimination is based on the criteria put forward by Ebel (1972) cited in Ovwigho (2013).

## FINDING ANG DISCUSSION

### Findings

In this study, students' responses to the two selected tests which has been scored by the English teacher assigned to teach English in the ninth grade were analyzed in order to determine whether the tests were good in terms of their validity, reliability, item facility and item discrimination.

### Validity of Summative Test of the Fifth Semester Examination 2019/2020

As pointed out earlier, the validity of summative test used in the fifth summative evaluation 2020/2021 was analyzed using Point Biserial correlation. This statistic was used because the test analyzed is an objective type test in multiple-choice format. Results of the analysis indicate that of 40 multiple-choice items analyzed. All of these items are valid with validity coefficients ranged between 0.53 and 0.98. Usually, an item with minimally 0.3 is considered valid.

### Reliability of Fifth semester summative test 2020/2021

Based on the data mentioned in Appendix A, reliability of the test was analyzed using KR-20. The statistic is commonly used to find out if an objective type test is reliable or not. For this purpose, the sums of X, $X^2$, pq, and $s^2$ were calculated. The results are presented in Table 3.

Table 3. The sums of X, $X^2$, pq, and $s^2$

| | |
|---|---|
| **$\sum X$** | **802** |
| **$\sum X^2$** | 21066 |
| **$\sum pq$** | 9.42 |
| **$S^2$** | 47.7 |

Based on the results mentioned in Table 3, statistical analysis using KR-20 was carried out. Result of the analysis indicated that $r_{(KR-20)}$ was 0.82. An item with coefficient larger than 0.6 is considered reliable Thus, it can be said that the test is highly reliable.

### Item Facility of Fifth semester summative test 2020/2021

As with the analysis of validity and reliability, item facility of fifth semester summative test 2020/2021 was analyzed based on the data obtained. The results of the analysis were then matched with the criteria for acceptable item facility set by Musa et al (2019, p. 1478) as mentioned in Table 4.

Table 4. Criteria for acceptable Item Facility

| Item facility index (p) | No. items | Item evaluation |
|---|---|---|
| **≥ (above) 0.70** | - | too easy item |
| **0.30 - 0.70** | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, & 40 | Recommended, good or acceptable |
| **≤ (below) 0.30** | - | too difficult item |

Results of item facility analysis show that item facility indexes range between the lowest 0.55 and the highest 0.67. This indicates that, based on item facility criteria. All the items in the test were considered good or acceptable.

### 4.1.4 Item Discrimination of Fifth Semester Summative Test 2020/2021

Similar as the analysis of item facility, the analysis of item discrimination index was based on the data enclosed in Appendix A. Item discrimination was analyzed using KR-20, the statistic used for analyzing item discrimination of objective type tests. The Discrimination Indexes of Fifth Semester Summative Test 2020/2021 later on were matched with the item discrimination criteria put forward by Ebel (1972 cited in Ovwigho, 2013).

Table 5. Evaluation of Item discrimination of Fifth Semester Summative Test

| Index of discrimination | No. items | Item evaluation |
|---|---|---|
| **0.40 and above** | 1, 2, 3, 4, 5, 8, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 26, 29, 32, 35, 37, 39, & 40 | Recommended, Very good items; accept |
| **0.30 - 0.39** | 6, 7, 10, 11, 24, 27, 28, 30, 31, 33, 34, 36, & 38 | Reasonably good |
| **0.20 - 0.29** | - | Marginal item usually subject to improvement |
| **Below 0.19** | - | Poor items to be rejected / improved by revision |

It can be said that of 40 multiple-choice items in the test, 27 items were considered very good items, and 13 items were reasonably good.

### Validity of Summative Test Used in the Sixth Semester Examination 2019/2020

As with analysis of validity of the fifth semester test, the validity of summative test used in the sixth summative evaluation 2019/2020 was analyzed using the same statistic technique for objective type tests. The statistical analysis was conducted and the results showed that the lowest validity coefficient is 0.53 (item no. 24) and the highest 0.85 (item no. 13). It can be concluded that the test is good in its validity.

### Reliability of Summative Test of the Sixth Semester Examination 2019/2020

Based on the data obtained, reliability of the test was analyzed using the same statistical technique as in the analysis of the sixth semester test. For this purpose, the sums of X, $X^2$, pq, and $s^2$ were calculated. The results are presented below.

Table 6. The sums of X, $X^2$, pq, and $s^2$

| | |
|---|---|
| **$\sum X$** | **993** |
| **$\sum X^2$** | 33315 |
| **$\sum pq$** | 11.5 |
| **$S^2$** | 46.1 |

Based on the criterion, it can then be said that the test is highly reliable

### Item Facility of Sixth Semester Summative Test 2019/2020

Similar with the analysis of the fifth semester summative test, item facility analysis of the sixth semester test was done based on the data enclosed hereby in Appendix B. Results of the analysis are presented in Table 4I. The item facility indexes were then matched with item facility criteria. The results are shown in Table 7.

Table 7 Evaluation of Item Facility of Sixth Semester Summative Test

| Item facility index (p) | No. items | Item evaluation |
|---|---|---|
| **≥ (above) 0.70** | - | too easy item |
| **0.30 - 0.70** | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, | Recommended, good or acceptable |

| | | |
|---|---|---|
| | 42, 43, 44, 45, 46, 47, 48, 49, & 50. | |
| **≤ (below) 0.30** | - | too difficult item |

Based on the criteria just mentioned, it can be stated that all the items in the sixth summative test are considered good or acceptable in terms of their item facility indexes.

**4.2.4 Item Discrimination of Sixth Semester Summative Test 2019/2020**

As the analysis of item discrimination of the fifth semester test, the item discrimination analysis of the sixth summative test was based on the obtained data were analyzed and results of the analysis were matched with the criteria put forward by Ebel (1972 (cited in Ovwigho, 2013).

Table 8. Evaluation of Item discrimination of Fifth Semester Summative Test

| **Discrimination Index** | **No. items** | **Item evaluation** |
|---|---|---|
| **0.40 and above** | 1, 3, 4, 5, 6, 7, 8, 9, 10, 13, 18, 22, 24, 25, 30, 31, 32, 33, 35, 36, 37, 38, 39, 40, 41, 44, 46, 47, 48, 49, 50. | Recommended, Very good items; accept (30) |
| **0.30 - 0.39** | 2, 11, 12, 14, 15, 16, 17, 19, 20, 21, 23, 26, 27, 28, 34, 42, 43, 45, & 46 | Reasonably good (19) |
| **0.20 - 0.29** | 29 | Marginal item usually subject to improvement |
| **Below 0.19** | - | Poor items to be rejected / improved by revision |

Based on the criteria in Table 8, it can then be said that of fifty items in all, 30 are very good or acceptable, and, therefore, are recommended for use, 19 are reasonably good, and only one is marginal item that subject to improvement.

Results of the data analysis of the two summative tests can be summarized as follows:

Concerning the fifth summative test, the following results are obtained: (1) item validity of the test ranges between 0.53 and 0.98 meaning that the test is good in its validity; (2) reliability analysis shows that rKR-20 is 0.82, meaning that the test is highly reliable; (3) item facility index ranges between 0.55 and 0.67,meaning that the test is good in terms of its item facility index,; and (4) of 40 multiple-choice items in the test, 27 items were considered very good items, and 13 items were reasonably good, meaning that the test is good in terms of its discrimination power. .

Concerning the sixth semester summative test, the following results are obtained: (1) the lowest validity coefficient is 0.53 (item no. 24) and the highest 0.85 (item no. 13); it can be concluded that the test is good in its validity; (2) reliability analysis shows that $r_{KR-20}$ is 0.88, meaning that the test is highly reliable; (3) all the items in the sixth summative test are considered good or acceptable in terms of their item facility indexes.; and (4) of fifty items in all, 30 are very good or acceptable, and, therefore, are recommended for use, 19 are reasonably good, and only one is marginal item that subject to improvement. It can be concluded that since almost all the

items are good in their discrimination power, the test is considered good in discriminating those who studied and those who did not.

**Discussion**

Evaluation of students' progress at school relies, among others, on test of various kinds, students' active involvement in the teaching and learning process, their home assignments, attendance, etc. Of these, test plays the most important role. Tests used at school are generally used to obtain information about students' mastery of taught, and to determine whether the students are allowed to proceed to the next grade. Therefore, they must be they must be well-developed. a good test helps teachers make accurate decisions related to their students' progress.

The two summative tests being studied in the present study are well-developed because in general they are valid, reliable, and appropriate in their item facility and item discrimination. This finding contradicts with the finding of the survey conducted by Direktorat Pembinaan Sekolah Menengah, Kementerian Pendidikan Nasional, 2010, p.105) which revealed that:

> Penilaian hasil belajar peserta didik yang dilakukan pendidik ternyata belum sepenuhnya menggambarkan tingkat pencapaian kompetensi peserta didik yang sesungguhnya, karena guru (1) tidak membuat kisi-kisi dalam pengembangan butir soal; (2) dalam membuat soal tidak mengikuti kaidah-kaidah penulisan soal yang baik dan benar, sehingga hasil belajar peserta didik belum menggambarkan kompetensi yang dituntut, (3) belum membuat soal secara mandiri (hanya mencontoh, mengcopy contoh-contoh soal dari guru lain atau dari buku Lembar Kerja Siswa yang dijual di pasaran) dan (4) tidak melakukan analisis butir soal, sehingga tidak mengetahui indicator/KD (kompetensi dasar) mana yang belum mampu dicapai peserta didik. Kondisi ini disebabkan karena guru belum memahami dan belum mengembangkan soal, dan menganalisis butir soal sesuai prinsip, mekanisme, dan prosedur penilaian (2010:105).

Recent studies conducted by Badara (2016), Rohmah (2018), Setiabudi et al., (2019) and Mulyani (2020) revealed similar results as that conducted by Direktorat Pembinaan Sekolah Menengah. Badara (2016) found that the tests made by Indonesia language teachers are the teacher-made test invalid and unreliable. Mulyani (2020) also found that the test used was not valid and reliable. Setiabudi et al (2019) and Rohjmah (2018) revealed found that test items were not appropriate in their item facility, item discrimination and effectiveness of distractors. This all means that teachers at junior and senior are weak in test development. Consequently, the test commonly used at junior and senior high schools do not effectively assess students' learning outcomes.

The previous research findings indicate that teachers, including English teachers, should be given more training in test development. Such a training is crucial because test results contribute to the decisions teacher made concerning their students' progress at school. In addition, it is important for teachers not to rely only on test in making their decisions. Other factors such as student active involvement in teaching and learning process and home assignments should also be considered. In this way, we can avoid using test scores to punish the students.

## CONCLUSION AND SUGGESTION

Results of the data analysis and the discussion above finally lead the researcher to derive the following conclusions.

1. Concerning the validity and reliability of *the fifth semester 2020/2021 summative test*, the data analysis shows that all the items are valid with validity coefficient ranges between 0.53 and 0.98, whereas reliability coefficient is 0.82, meaning that the test is highly reliable. It can be concluded that the test is valid and highly reliable.
2. Concerning validity and reliability of *the sixth semester summative test* in 2019/2020, the analysis indicates that all the items are valid with validity coefficient ranges between 0.53 and 0.85, whereas reliability coefficient is 0.88, meaning that the test is highly reliable. It can be concluded that the test is good in its validity and highly reliable.
3. Item analysis item facility of the fifth semester 2020/2021 summative test shows that item facility index of the 50 items in the test ranges between 0.55 and 0.67, meaning that the test is good in terms of its item facility index; and of 40 multiple-choice items in the sixth semester 22019/2020 summative test, 27 items are considered very good items, and 13 items are reasonably good in their discrimination power. It can be concluded that the test is good in terms of its item facility and discrimination power.
4. Item analysis of the sixth semester 2019/2020 summative test shows that in terms of item facility, all the 50 items are recommended for use. In terms of item discrimination, the analysis shows that 30 are very good or acceptable, and, therefore, are recommended for use, 19 are reasonably good, and only 1 is marginal and subject to improvement. It can be concluded that the test is good in its item facility and discrimination power.

In general, the two summative tests, one used in the fifth semester 2020 and the other one in the sixth semester 2019 are good in terms of the characteristics of a good objective type test in multiple-choice format.

Based on this finding, it is necessary for the researcher to put forward these suggestions:

1. Since teachers are weak in test development, it is important for the government to provide in-service training for teachers focusing on test development. In this way, test quality can be increased.
2. No test is perfect. Therefore, teachers, including English teachers, are expected not to base their decision concerning students' progress based only on test results. Other factors such as active involvement in teaching and learning process, home assignment and test administration should also be taken into consideration.
3. Lots of tests of various type are used at school. Therefore, more studies are still needed in order to obtain more information about the quality of tests used at schools.

## REFERENCES


Alderson, J. C. & Hughes, A. 1981. *Issues in language testing.* London: British Council.

Badara, Aris. 2016. *The Quality of the Indonesian Language Teacher-Made Tests at Junior School Level*. Prosiding ICTTEFKIP UNS 2015. Vol 1, No. 1.

Bachman, L. F., & Palmer, A. 1996. *Language testing in practice*. Oxford: Oxford University Press.

Brown, James Dean. 2005. *Testing in Language Programs: A comprehensive Guide to English Language Assessment* New York: McGraw-Hill EL/ELT.

Brown, H. Douglas. 2001. Teaching by Principle and Interactive Approach

to language pedagogy. New York: Longman Inc.

Direktorat Pembinaan Sekolah Menengah, Kementerian Pendidikan Nasional 2010. *Supervision and Evaluasi Implementation of KTSP 2009*. Jakarta: DIKNAS.

Ebel, R. L. 1979. *Essential of Educational Measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Gay, L. R., Mills, Geoffrey E., and Airasian, Peter. 2012. *Educational Research: Competencies for Analysis and Applications*. Pearson Education, Inc.

Hadi, Sutrisno. 1997. *Metodologi Research*. Yogyakarta: Andi Offset.

Hamed Taherdoost, Hamed. 2016. *Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research.* IJARM, Vol. 5, No. 3, Page: 28-36, ISSN: 2296-1747

He Lianzhen & LvZhouyang. 2013. *A New Perspective of Language Testing Research*: *Critical Language Testing*. 43(6): 164-173.

Heaton, J.B. 1975. *Writing English Language Test.* London: Longman Limited.

**Korb. Calculating Reliability of Quantitative Measures. http://korbedpsych.com/LinkedFiles/CalculatingReliability.pdf**

Kubai, Edwin. 2019. *Reliability and Validity of Research Instruments*. Conference: NMK conference. Project: Critical Analysis of policies on Special Education in Kenya. kubaiedwin@yahoo.com

Lalogiroth, A., & Tatipang, D. P. (2020). *An Analysis of English National Exam and English Teachers'perception Using Bloom's Revised Taxonomy*. Journal of English Culture, Language, Literature and Education, 8(1), 1-21.

Lebagi, Desrin., Sumardi, Sumardi., & 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.

Lumentut, Y., & Lengkoan, F. (2021). *The Relationships of Psycholinguistics in Acquisition and Language Learning*. Journal of English Culture, Language, Literature and Education, 9(1), 17-26.

Mulyani, Heni., Tanuatmodjo, Heraeni & Iskandar, Rangga.. 2020. *Quality analysis of teacher-made tests in financial accounting subject at vocational high schools*. Jurnal Pendidikan Vokasi, Vol 10, No 1.

Oller, John W. 1979. *Language Tests at School*. London: Longman.

Oluwatayo, J. 2012. *Validity and reliability issues in educational research*. Journal of Educational and Social Research 2, 391-400.

Ovwigho, B. O. 2013. *Empirical Demonstration of Techniques for Computing the Discrimination Power of a Dichotomous Item Response Test*. IOSR Journal of Research and Method in Education; 3(2): 12-17.

Pikirang, C. C., Liando, N., & Wuntu, C. N. (2021). *A Correlational Study Between Learners'satisfactions With Offline Class and English Self-Efficacy During The Covid-19 Pandemic*. Journal of English Culture, Language, Literature and Education, 9(1), 73-85.

**Popham, W. James. 2009. *All About Assessment/Diagnosing the Diagnostic Test*. Educational Leadership, Vol. 66, No. 6, pp. 90-91.**

Primadani, Arin Eka & Sulistyo, Gunadi Harry. 2014. *An Analysis of a Midterm English Test of the Seven Grade Accelerated Class at SMPN 3 Malang*. Email:

arina.gee2@gmail.com & emailgun.now@gmail.com

Rajhy, Hussein Ahmed Abdo. 2014. *Five Characteristics of a good Language Test*. National Journal of Extensive Education and Interdisciplinary Research, Volume II, Issue IV, p. 61-66. ISSN: 2320-1460. ISSN: 2320-1460.

Rohmah, Naelul. 2018. *Validity and Reliability Study on Teacher-Made Assessment for English Mid-Term Examination*. Advances in Social Science, Education and Humanities Research, volume 254. Eleventh Conference on Applied Linguistics.

Saefurrohman & Balinas, Elvira S. 2016. *English Teachers Classroom Assessment*. English Department Faculty of Languages and Arts, Semarang State University.

Setiabudi, Agung., Mulyadi, Mulyadi., & Puspita, Hilda. 2019. *An Analysis of Validity and Reliability of a Teacher-Made Test.* Journal of English Education and Teaching, Vol 3, No 4.

Straub, D., & Boudreau, M.-C. & Gefen, D. 2004. *Validation guidelines for IS positivist research*. Communications of the Association for Information Systems, 13, 380-427.

Yohana Putri. 2009. *An Analysis of Teacher-Made English Final Second Semester Test for the Year Eleven Students of SMAN 1 Ambarawa in the Academic Year of 2008/2009 Based on the Representativeness of Content Standard*. English Department Faculty of Languages and Arts, Semarang State University.

Weiss, C.H. 1972. *Evaluation Research*: *Methods of Assessing Program Effectiveness*. Englewood Cliffs (NJ), USA: Prentice-Hall.

Wiersma, William & Jurs, Stephen G.1990. *Educational Measurement and Testing*, second edition, Boston: Allyn and Bacon.