THESIS

AN ANALYSIS OF THE ENGLISH TEST FOR SECOND GRADE STUDENTS AT SMP MUHAMMADIYAH MAMUJU: THE TEACHER'S DEVIATION IN CONSTRUCTING ENGLISH TEST

SUNDARI
H0120318

The Undergraduate Thesis was Written and Submitted in a Part-Fulfillment of the Requirements for undergraduate Thesis Degree Education

ENGLISH EDUCATION STUDY PROGRAM
FACULTY OF TEACHER TRAINING AND EDUCATION
UNIVERSITAS SULAWESI BARAT
2025

APPROVAL SHEET

AN ANALYSIS OF THE ENGLISH TEST FOR SECOND GRADE STUDENTS AT SMP MUHAMMADIYAH MAMUJU: THE TEACHER'S DEVIATION IN CONSTRUCTING ENGLISH TEST

SUNDARI

H0120318

Has been successfully defended the thesis in front of the Examiner Team of Faculty of Teacher Training and Education on June 24th 2025

EXAMINATION

Chair of the Examiner	: Prof. Dr. H. Ruslan, M.Pd.	(.....................)
Secretary of Exam Committee	: Hustiana, S.Pd., M.Pd.	(.....................)
Supervisor I	: Amrang, S.Pd., M.Pd.	(.....................)
Supervisor II	: Dr. Umar, S.Pd., M.Pd.	(.....................)
Examiner I	: Ridwan, S.Pd.I., M.Pd.	(.....................)
Examiner II	: Nurul Imansari, S.S., M.A.	(.....................)

Majene, June 24th 2025

FACULTY OF TEACHER TRAINING AND EDUCATION
UNIVERSITAS SULAWESI BARAT
Dean,

Prof. Dr. H. RUSLAN, M.Pd.
NIP. 196312311990031028

STATEMENT OF WORK ORIGINALITY

The one who is filed below:

Student's Name	: Sundari
NIM	: H0120318
Study Program	: English Education Study Program

Hereby states that the thesis Sundari original work and has never been submitted for an undergraduate degree in a higher school, and as far I am concern in this thesis no work or opinion has been written or published by others expect has been referred explicity in this document and listed in the bibliography.

If in the future, it is proven that this thesis is a copy, I am willing to accept the sanction for my act.

Majene, 10 February 2025

Signed by

Sundari

H0120318

ABSTRAK

SUNDARI: Menganalisis Tes Bahasa Inggris untuk Siswa Kelas Dua di Smp Muhammadiyah Mamuju: Deviasi Guru dalam Menyusun Tes Bahasa Inggris. **Skripsi, Majene: Fakultas Keguruan Dan Ilmu Pendidikan, Universitas Sulawesi Barat, 2025**

Penelitian ini bertujuan untuk menganalisis kualitas butir soal pilihan ganda mata pelajaran Bahasa Inggris serta mengidentifikasi faktor-faktor kesulitan yang dihadapi guru dalam penyusunan tes bahasa inggris di SMP Muhammadiyah Mamuju. Menggunakan metode campuran (mixed methods), penelitian ini melibatkan 33 siswa kelas VIII-1 dan satu orang guru. Sebanyak 20 butir soal dianalisis secara kuantitatif dan dilengkapi dengan data kualitatif. Hasil menunjukkan bahwa kualitas soal secara umum masih rendah; hanya 11 soal yang valid dan 9 tidak valid, reliabilitas tes sebesar 0,51, serta variasi tingkat kesulitan ada 12 kategori mudah, 7 moderate, dan ada 1 kategori sulit. Daya pembeda yang 9 kategori baik, 7 cukup dan ada 4 kategori buruk. Serta distractor yang bervariasi. Standar deviasi sebesar 2,8 menunjukkan kemampuan siswa yang cenderung homogen. Temuan kualitatif mengungkap ada enam kendala yang dihadapi guru, yaitu: (1) kesulitan membuat pengecoh yang efektif, (2) penentuan tingkat kesulitan soal yang subjektif, (3) tidak adanya analisis soal sebelum atau sesudah tes digunakan, (4) belum terjaminnya konsistensi dan keadilan soal, (5) soal belum mencakup seluruh level kognitif, serta (6) kurangnya penyesuaian soal dengan kemampuan siswa. Penelitian ini menyimpulkan perlunya pelatihan berkelanjutan dan dukungan teknis dalam pengembangan instrumen penilaian yang valid, reliabel, dan adil.

Kata kunci: Analisis tes, Standard deviasi.

ABSTRACT

SUNDARI: An Analysis of the English Test for Second Grade Students at SMP Muhammadiyah Mamuju: the Teacher's Deviation in Constructing English Test. **Thesis, Majene: Faculty of Teacher Training and Education, Universitas Sulawesi Barat, 2025**

This study aims to analyze the quality of multiple-choice items in English subjects and identify the difficulty factors faced by teachers in preparing English tests at SMP Muhammadiyah Mamuju. Using mixed methods, this study involved 33 students of class VIII-1 and one teacher. A total of 20 items were analyzed quantitatively and supplemented with qualitative data. The results showed that the quality of the questions was generally low; only 11 questions were valid and 9 were invalid, the reliability of the test was 0.51, and the variation in difficulty level was 12 easy categories, 7 moderate, and there was 1 difficult category. The differentiating power is 9 good categories, 7 satisfactory and there are 4 worst categories. As well as distractors that vary. The standard deviation of 2.8 indicates that students' abilities tend to be homogeneous. Qualitative findings revealed six obstacles faced by teachers, namely: (1) difficulty in making effective exemptions, (2) subjective determination of the level of difficulty of the questions, (3) no analysis of the questions before or after the test is used, (4) the consistency and fairness of the questions have not been guaranteed, (5) the questions have not covered all cognitive levels, and (6) the lack of adjustment of questions to students' abilities. This study concludes the need for continuous training and technical support in the development of valid, reliable, and fair assessment instruments.

Keywords: Test analysis, Standard deviation

CHAPTER I
INTRODUCTION

A. Background

A test is an instrument used to assess the abilities, skills, qualities, or knowledge of a group or individual based on measurable knowledge standards. That is, tests are designed to assess the extent to which a person or group of people has mastered certain material. A test can be a series of questions, tasks, or activities intended to collect data about a person's competence or understanding of a subject (Adom et al, 2020).

According to Brown (2004), a test is a method for measuring a person's ability, knowledge, or performance in a particular field. In the context of formal education in schools, tests are used to assess students' abilities or knowledge, place them at an appropriate level, evaluate learning progress, and provide feedback as part of the learning process. Brown said that tests are divided into two forms, namely teacher-made tests and standardized tests. A teacher-made test is designed by the teacher based on the curriculum and the lesson plan that have been applied during the lesson. It is intended to measure the success rate of students in achieving the target of the curriculum after the teaching learning process is done by the teacher. Therefore, the teacher must make logical and rational questions about what items are worth asking. This test is usually used for daily, formative, and general tests. While Standardized Tests are designed by experts such as teachers or expert institutions in the field of study. The test is standardized, that is, management is carried out based on standards and assumptions of uniform conditions so that the results of the assessment can be compared for different classes or schools.

According to Badan Pengembangan dan Pembinaan Bahasa, the word "deviation" means to disagree, go astray, or deviate from a rule. Thus, deviant behavior includes any form of action that goes against the standards or expectations set by the society or group. Deviance also might be experienced by teachers in educational field. Teacher deviance in creating tests has several important implications that can affect student learning and evaluation outcomes. Deviations can lead to unfairness in grading, where some students may gain an unfair advantage or disadvantage, deviations can affect the validity and reliability of the test, so that the results do not accurately reflect students' abilities, tests that are perceived as unfair or irrelevant can demotivate students to learn, and deviations can be an indicator that teachers need more training in test design and scoring.

Reducing deviations in test creation requires careful planning, a good understanding of assessment principles, and awareness of potential biases that may arise. Continuous training and professional development for teachers is also crucial to ensure fair and effective tests.

According to National Center for Educational Statistics (2019), Deviations in test preparation are errors or practices that are not in accordance with good principles and guidelines in test development. As a result in tests that are invalid, unreliable, and unfair to students. This can result in tests that are invalid, unreliable, and unfair to students. there are several examples of deviations that are often made by teachers in preparing tests, such as errors in compiling items, errors in carrying out tests and errors in assessing test results. the impact is that students do not get accurate information about their abilities, Teachers lose the trust of students and parents, Students feel frustrated and unmotivated and Schools lose their reputation.

In fact, teacher rarely does trials on the questions to be used, including analyzing the quality of each item to be tested, so most of them cannot be identified as proper tests. This is due to the lack of teacher time and teacher understanding. Even though analyzing the items is an activity that must be carried out by a teacher.

According to Nurgiyantoro (2010), the reason why the analysis of each item must be carried out is because it will produce quality questions in subsequent tests and to find out what are the strengths and weaknesses of the previous items so that the items can be selected, revised, and can immediately know the problems in the items and will be an indication in the next test. If the tests made by the teacher are not in accordance with what has been determined, it will have an adverse effect on students; namely, the interests, talents, and understanding of students cannot be measured, so the teacher cannot classify students with different ability levels.

This research will be carried out at SMP Muhammadiyah Mamuju, which has elementary, junior high and senior high school education units and is an integral

Boarding Schools (integrated with schools and madrasah) with the name of the pesantren Muhammadiyah Boarding School At-Tanwir. Is located on Jl. Soekarno Hatta No 35 Mamuju, Karema District, Mamuju Subdistrict, Mamuju Regency, West Sulawesi

The problem that research found in choosing SMP Muhammadiyah Mamuju, based on the results of preliminary observations that showed the lack of teachers' knowledge in analyzing the tests they made, as well as students' difficulties in understanding the tests. The initial observations suggest that teachers may not fully understand the assessment criteria or how to analyze their own test results. In addition, students may have difficulty in understanding the instructions and content of the test.

background

According to Arikunto (2013), a good test item must meet several criteria, such as validity, reliability, level of difficulty, discriminating power, and effectiveness of distractors. However, in practice, many teachers still do not understand or apply these principles comprehensively. This results in test items that do not align with the learning indicators, are too easy or too difficult, and are unable to distinguish between students who truly master the material and those who do not. Sudijono (2011) explains that test items that do not meet these criteria can give a misleading picture of students' abilities. In addition, various obstacles are faced by teachers in preparing questions, such as time constraints and lack of training. These factors often lead to deviations in question preparation, such as the use of confusing language, illogical answer choices, or the absence of a validation and reliability checking process for the questions.

Moreover, statistical aspects such as standard deviation, which can be used to assess the distribution of students' scores, are often not considered in the preparation or analysis of test results. Besides deviations in the content preparation of the questions, technical and statistical aspects such as validity, reliability, and standard deviation are also often overlooked. Many teachers do not analyze the questions that have been created, whether manually or using statistical tools. However, by conducting analyses such as difficulty levels, discrimination index, and standard deviation, teachers can determine whether the questions the teacher has formulated are appropriate and fair for all students.

Based on this issue, it is important to conduct a study that comprehensively examines the quality of multiple-choice questions prepared by teachers, particularly in English subjects at the junior high school level. This research focuses on three main aspects. First, analyzing multiple-choice questions based on technical aspects, namely validity, reliability, difficulty level, discrimination index, and distractors, to determine whether the questions created can accurately measure students' abilities. Second, tracing and identifying the factors that cause teachers to make mistakes or deviations in crafting questions. Third, to analyze the implementation of standard deviation in test construction to see how far teachers consider the distribution of student scores as an indicator of the test quality. The researcher will carry out research with the title " An Analysis of the English Test for Second Grade Students at SMP Muhammadiyah Mamuju: The Teacher's Deviation in Constructing English Test."

B. Problem Identification

Based on the background above, the researcher identified a problem from English teachers at SMP Muhammadiyah Mamuju, it is there are several forms of tests applied without proper analysis such as validation, reliability, level of difficulty, discrimination index, and distractors. It is important to find out the standard deviation of English tests.

C. Problem Limitation and Formulation

Based on the background above, the researcher came up with research formulation, they are:

1. How is the quality of multiple-choice English test made by teacher based on validity, reliability, level of difficulty, discrimination index, and distractors at SMP Muhammadiyah Mamuju?
2. What are the factors that cause English teachers' difficulty in creating multiple-choice English tests for second-grade students at SMP Muhammadiyah Mamuju?

3. How is the standard deviation of the English tests at SMP Muhammadiyah Mamuju?

D. Objective of Research

Based on the researcher's formulation of the problem previously outlined in the report, this study aims to identify of following objectives:

1. To know the quality of multiple-choice English test made by teachers based on validity, reliability, difficulty of level, discrimination index, and distractors at SMP Muhammadiyah Mamuju.
2. To know the factors that cause English teachers' difficulty in creating multiple-choice English tests for second grade students at SMP Muhammadiyah Mamuju
3. To know the standard deviation the English tests at SMP Muhammadiyah Mamuju.

E. Research Benefits

1. The research is expected to teachers on determining the quality of tests made by them with the level of ability obtained by students. whether the test is maximized for students or has the quality to improve educational characteristics.
2. This research hopefully can help students in creating a supportive and effective learning environment, helping students achieve their academic potential to the maximum according to their ability level.
3. This research can give a deep understanding of various aspects of educational evaluation, especially in test making.

CHAPTER II
LITERATURE REVIEW

A. Previous Related Studies

There are many research findings which are related to this research, some of previous findings are described below:

The first reached conducted by Indrayani, et al (2020), with the title "The Analysis of the Teacher-Made Multiple-Choice Tests Quality for English Subject" The study analyzed the quality of teacher-made multiple-choice English tests used for summative assessments using a descriptive research design and document analysis. The data was compared against 18 established rules for good multiple-choice test preparation. The results were tabulated and categorized into five quality levels: very good, good, sufficient, poor, and very poor. The study assessed English tests for grades VII, VIII, and IX, consisting of 20, 30, and 30 questions respectively, each with four answer options. The results showed that 99% of the items were of very good quality and 1% were of good quality. However, issues with punctuation and capitalization were common, with only 71% compliance. To address these problems, the study recommends peer reviews, double-checking, and editing by teachers. Compliance with other rules varied between 84% and 100%.

The similarity of the above research with this research both analyzing the quality of multiple-choice tests, while the difference from this research is that the research methods are different and the data collection process is that the research mentioned above analyzed data by comparing 18 rules made by Nurkencana while this research does not use any theory only focuses on analyzing tests based on validity, reliability, level of difficulty, discrimination index and distractors.

The second research, written by Hartati & Yogi (2019), titled "Item Analysis for a Better Quality Test," focuses on evaluating the quality of tests created by English teachers. This study investigates the quality of multiple-choice items regarding difficulty level, discriminative power, and the effectiveness of distractors. Qualitative methods are employed using summative English tests and student answer sheets. The quality of the summative English test items is assessed through qualitative analysis, while simple quantitative analysis is conducted to analyze facility value (FV), differentiating power (DP), and distractor effectiveness. The research involved a total of 65 students at SMA Muhammadiyah in the first semester.

The similarity between this study and the previous study is that both aim to provide recommendations for test improvement to be more effective in measuring students' abilities. This study uses a mixed method that combines qualitative and quantitative approaches. and analyzes English tests for second grade students and evaluates standard deviations in compiling English tests and assesses the quality of multiple-choice tests in terms of validity, reliability, difficulty level, distinguishing power and exemplars, while the previous study used a cross-sectional study to evaluate MCQs. The previous study only assessed the quality of multiple-choice questions in terms of difficulty level, differentiating power, and efficiency of exemplars.

The next research is research from Darmawan et al (2022), "A Test-Items Analysis of English Teacher-Made Test". This study aims to showed moderate difficulty, good item discrimination and high reliability, with 60% of the questions proving valid. This study used two types of instruments to analyze the quality of tests created by English teachers at SMAN 8 Pontianak. First, they used the Master Tap application to analyze the structure and characteristics of the test questions. Second, they used SPSS version 16 software to conduct statistical analysis of the data collected from the tests. With the combination of these two instruments, the research was able to comprehensively investigate the validity, reliability, difficulty level, item discrimination and distractors of the test.

The similarities between these researchers are analyzing tests based on validity, reliability, level of difficulty, and distributors, the previous researcher used quantitative descriptive analysis with data obtained from English teacher-made tests, consisting of 40 items with 257 test takers of twelfth grade students. The researcher analyzed the test questions using a combination of the Master TAP application and SPSS Version 16. while this researcher used a mixed method and found out the teacher's deviations in making tests.

Research according to Marsevani, M. (2022), Entitled "Item analysis of multiple-choice questions: An assessment of young learners" this study aims to determine the quality of multiple-choice tests in public elementary schools based on difficult level, discrimination power, and distractor efficiency. The study used cross-sectional in obtaining information and evaluating student tests. In the middle of term, the students who were given 10 multiple choice tests were grade 5 students totaling 40 students. The time allotted was 60 minutes. Each item was analyzed using ANATES Ver. 4.0.9 and SPSS Ver. The analysis showed most of the multiple-choice questions had an appropriate level of difficulty, with 80% meeting the standard, while 20% were too easy. The majority of questions had good differentiating power. However, there were two questions with ineffective answer choices.

The similarities between the two studies focus on the analysis of teacher-made English tests, both of which analyze aspects of items such as difficulty level, differentiating power, validity, and reliability. This study uses a mixed method, which combines quantitative and qualitative analysis, while the previous study used descriptive quantitative analysis.

A study by Anggriani, E. (2021), entitled "An Analysis Of Multiple-Choice Items Of English Final Semester Test Made By English Teacher" aims to determine the quality of teacher-made test items based on the level of difficulty and differentiating power at MA AL-Muthmainnah grade 10 school in the final semester exam. The study used quantitative methods using descriptive analysis. From the results of the study of 25 items on the English end-of-semester exam for tenth grade, it was found that 68% of the items had a medium difficulty level, 24% had a high difficulty level, and 8% had a low difficulty level. Regarding differentiating power, 32% of the questions had low differentiating power, 44% had sufficient differentiating power, and 24% had good differentiating power.

The similarities between this study and the previous one both use documentation studies as a data collection technique and involve evaluating the quality criteria of exam questions, although the specific aspects evaluated may differ. The previous research used a quantitative approach with descriptive and statistical analysis. This research uses a mixed method, which means you will combine quantitative and qualitative approaches. The previous study aimed to determine the level of difficulty and differentiating power of end-of-semester exam items. while this study uses validity, reliability, level of difficulty, discrimination index, and distractors.

Based on the previous related findings there are some differences with this research. There are similarities in analyzing the quality of multiple-choice tests using descriptive research. However, there are differences in analysis methods, research subjects, and research approaches. In addition, there are differences in the variables analyzed and the way of approaching data analysis, thus contributing additionally to the understanding of the validity, reliability, and effectiveness of distractors in teacher-made tests. These differences are important for broadening the scope of research on multiple-choice test evaluation, given the variety of different contexts and research objectives.

The most common error factors found in all studies were inappropriate item difficulty, low item discriminating power, and ineffective distractors. The five studies consistently show that many questions are too easy or too difficult, questions that fail to differentiate between high and low ability students, and wrong answer options (distractors) that do not function properly in tricking students.

In addition, a low content validity factor was also found in most studies (Indrayani et al., Darmawan et al., Marsevani, and Anggriani). This indicates that some questions do not match the basic competencies or materials that have been taught to students. Another common error is in question construction, such as the use of inappropriate language, unclear question stems, and inhomogeneous answer options. All studies noted deficiencies in this aspect.

In some studies, low test reliability was also found, especially in studies by Hartati & Yogi (2019) and Darmawan et al. (2022). The low reliability indicates that the test results given by teachers are not consistent enough to be used as a reliable evaluation tool. In addition, two studies (Indrayani et al. and Marsevani) highlighted inconsistent question writing formats, such as non-uniformity in question layout and answer options. Unclear question instructions were also noted in Darmawan et al. and Marsevani's studies, which showed that some questions did not provide sufficient instructions to students.

B. Theoretical Framework

1. Teacher's Difficulties in Constructing a Test

In learning activities, teachers are not only tasked with delivering material, assessing student learning outcomes, including in the form of tests. The test preparation process is an important stage that requires a deep understanding of the formulation of learning objectives, and adjustments to students' thinking abilities. However, in reality, many teachers experience obstacles in every stage of test preparation, from planning, writing questions, to choosing the right test form so that the resulting test is truly valid, reliable, and follows learning needs.

According to Brown (2004), there are several factors that can cause difficulties for teachers in developing effective and quality tests. Some of the main challenges often faced by teachers in this process include:

a. Time Limitations

Teachers are often caught up in many other tasks and obligations, such as teaching, planning learning activities, carrying out administrative tasks, and managing the classroom. These time constraints often prevent them from developing tests carefully and in-depth. Limited time makes it difficult for teachers to develop questions that not only measure learning outcomes accurately, but also creative and varied questions in accordance with the learning objectives that have been set.

b. Limited Resources

Many teachers do not have sufficient access to the materials or references needed to create quality test questions. Limited resources, both in the form of question references, technological devices, and other supporting resources, greatly limit teachers' ability to design tests that can cover various aspects of student skills. This hampers teachers' creativity in producing tests that are not only up to standard but also interesting to students.

c. Lack of Professional Training or Education

Although test development is an important skill in the teaching profession, many teachers do not receive specialized professional training or education in this area. The lack of training in designing valid and reliable tests means that many teachers do not clearly understand how to create tests that can truly measure student achievement objectively. Without adequate training, they also struggle to understand the more technical aspects of assessment, such as item validity, consistency, and accuracy in choosing the right item types.

d. Difficulty in Measuring Different Cognitive Levels

Developing questions that are able to measure various levels of students' cognitive abilities is a big challenge. Teachers are often more comfortable with questions that measure basic skills such as remembering or understanding information. However, developing questions that can measure higher-order thinking skills, such as analysis, synthesis, and evaluation, in accordance with Bloom's taxonomy, requires deeper understanding and skills. This becomes more difficult if teachers are not trained in designing questions that develop students' critical thinking skills.

e. Compatibility of Questions with Learning Objectives

In the test development process, one of the most common difficulties faced by teachers is ensuring that the questions actually reflect the learning objectives to be achieved. Teachers often find it difficult to develop questions that not only measure students' basic knowledge, but also their ability to achieve more complex and holistic learning objectives. The match between test items and learning objectives is essential so that test results can accurately reflect students' abilities in terms of the skills that should be mastered.

These difficulties illustrate that while tests are an important part of learning evaluation, many factors can affect the quality of tests developed by teachers. Effective test development requires careful planning, practiced skills and sufficient support from available resources and training.

2. **Definition of Standard Deviation**

According to Sugiyono (2018), standard deviation is a measure of data distribution that shows how far the data varies from the average value. A small standard deviation indicates homogeneous data, while a large standard deviation indicates heterogeneous data.

Standard deviation is a statistical measure used to describe how much variation or spread a data set has from its average value. The smaller the standard deviation value, the closer the data is to the average (homogeneous), while the greater the standard deviation, the more the data is spread away from the average (heterogeneous).

Brown (2014 said, Standard deviation is a measure of how much test scores are dispersed from the mean. A low standard deviation indicates that test scores are more centered around the mean, whereas a high standard deviation indicates that test scores are more dispersed. Therefore, teachers should consider the purpose of the test, the context in which it is used, and the characteristics of the students when determining the appropriate standard deviation for their tests.

3. **Standard Deviation Criteria**

The interpretation of standard deviation in educational research can be classified as follows (Sudjana, 2005):

Table 2.1 Criterion of Standard Deviation

Standard Deviation Value	**Interpretation of Data Distribution**
Very small (close to 0)	Data is very homogeneous, values are close to the average.
Low (1-5)	Data is fairly homogeneous, with low variation.
Moderate (5-10)	The data has moderate variation.
Large (>10)	Data is heterogeneous, values are spread far from the mean.

(Sudjana, 2005)

In the context of a 0-100 rating scale, a standard deviation below 5 is considered low and indicates that test takers' results are relatively uniform. Conversely, a standard deviation above 10 indicates a large difference in ability between test takers.

4. **Definition of Test**

According to Arifin (2013), tests are a method used to conduct measurements involving questions, statements, or tasks that participants must complete to evaluate behavior or the quantity of something. The function of tests is as a measuring instrument used to identify or measure something according to established rules.

According to Haryanto (2016), a test is an instrument or assessment tool that is systematically designed to measure the level of student learning achievement after following a certain learning process, generally in the form of a series of tasks or questions presented in a numerical format with answers, namely true or false.

However, tests can also take the form of description questions or other forms that are relevant to the purpose of the assessment. The main purpose of the test is to obtain objective and accurate information regarding the extent to which students have mastered the subject matter that has been taught, as well as to identify areas where students may still require further guidance or reinforcement. The results of these tests can then be used by the teacher as feedback to improve and enhance the quality of the learning process going forward.

In the book "Evaluation and Learning: Theory and Practice," tests can be defined as a collection of questions or tasks planned to gather information about educational or psychological traits or attributes, where each question or task has a correct answer or criteria. Tests are a common type of assessment that usually consists of a series of questions administered at a specific time and under relatively similar conditions to all students. From this understanding, it can be seen that tests can provide an overview of how the intensity of a person's behavior is compared to other students or with certain benchmarks. Thus, learning outcome tests can be defined as systematic tools or procedures for measuring student learning outcomes.

Based on the opinions of the experts above regarding the definition of a test, researcher can conclude that a test is an assessment tool used to measure students' understanding, abilities, and skills through a series of systematically designed tasks or questions. The main purpose of the test is to provide objective and accurate values about student achievement in a particular material or skill, thus enabling teachers and educational institutions to make informed decisions regarding student learning and development. Thus, tests are not only an evaluation instrument, but also an important tool in the teaching and learning process to improve the overall quality of education.

5. Function of the Test

There are 3 (three) test functions in the world of education or training, according Djaali and Muljono (2008):

a. Tests are used to measure participants' learning achievements, assess the progress achieved after the teaching and learning process, evaluate the success of teaching or training programs, and determine the next steps to achieve goals that have not been achieved.

b. Tests can be a motivator in learning by providing feedback in the form of scores, which effectively increases the motivation and intensity of participants' competence in teaching and learning activities.

c. Tests can be used to improve the quality of learning by extending training programs that receive good scores to be implemented in the following year.

5. Types of Test

a. Types of test based on its function

1) Formative test

Formative tests are conducted during learning, usually after completing a section or topic. Based on the results, learning can continue smoothly if the material is well understood, or additional support can be provided on areas of lesser mastery, (Wahyuningsih, 2015).

Formative tests provide feedback to students and teachers during learning to gauge students' understanding of the material and deploy teachers' teaching methods. such as daily tests or mid-semester. It helps improve learning effectiveness and allows students to know the extent to which they have mastered the lessons learned.

2) Summative test

Summative tests are conducted when a teacher wants to know the latest progress of their students. This assessment aims to determine whether learners have achieved the competency standards that have been set. The results of this test are used to assign grades based on the students' level of learning achievement, which is then used as the final semester or national exam grade according to Widyaningsih (2012).

Summative tests are final evaluations that cover all the material that has been taught in one semester as in the end-of-semester final test or national exam. The purpose is to assess student mastery and provide information to the teacher to determine whether the student is worthy of advancing to the next level or not.

This research focuses on analyzing the English test used as a summative evaluation for grade one students. Summative tests are evaluations conducted at the end of a learning period to assess students' overall learning outcomes. These summative tests are designed to measure students' understanding and skills after completing a specific learning period. The purpose of this study is to identify errors or deviations that occur in the preparation of a summative English test for grade one students, to improve the validity and reliability of the test.

b. The Types of Test Items

1) Objective Test Multiple-Choice

The are several types of objective tests, namely multiple-choice tests, true/false tests, matching tests, and short answer tests. This research focuses only on multiple choice. Multiple choice tests consist of a description or notification of an incomplete understanding, and to complete it must choose one of several possible answers from several choices ranging from 3 to 5 choices that have been provided. Possible answers consist of one correct answer, namely the answer key, and several exceptions.

The objective test, also known as a short answer test, emphasizes participants to provide brief answers, even just by selecting specific codes representing provided alternative answers, such as marking crosses, circling, or darkening the chosen answer options (Nurgiyantoro, 2017)

Objective tests are a form of test created to measure knowledge, understanding, and skills in a neutral manner, where the correctness or incorrectness of the answer can be measured clearly and does not depend on the subjective judgment of the assessor. Typically, objective tests use multiple-choice, true-false, or similar formats, which allow for quick and consistent scoring.

6. **Characteristics of a Good Test**

The characteristics of a good test include several important requirements, namely validity, reliability, objectivity, practicality, and economy. Validity indicates how well the test measures what it is supposed to measure, while reliability indicates the consistency of test results. Objectivity ensures that test results are not influenced by the subjectivity of the test compiler or implementer, practicability assesses the ease of test implementation, and economy is related to the efficient use of funds and time. A good test also pays attention to the discrimination index, level of difficulty, and answer distribution patterns to assess the quality of the questions and support the improvement of the learning process of students (Wahyuningsih, 2015).

a. Validity

1) Definition of Validity

Validity is one of the important components in assessing learning outcomes, as it indicates the extent to which a test instrument can measure what it is supposed to measure. Arikunto (2013) explains that validity is a measure that indicates the accuracy level of a test regarding its measurement function. In other words, a question is considered valid if its content and form align with the learning objectives to be achieved. If the learning objective is to measure reading comprehension skills, then the questions formulated must truly assess that ability, not other abilities such as memorization or guessing. In line with this, Nitko and Brookhart (2014) emphasize that validity is not only related to the questions themselves but also to the extent to which the test results can be used to draw correct conclusions about student abilities.

In the book Learning Evaluation: Theory and Practice, validity refers to how accurately a test measures the aspects of teaching material or behavior that it is supposed to measure. However, the concept of validity is also related to the purpose of measurement, which refers to the accuracy of the test in producing data relevant to the purpose or decision to be made. Generally, there are three known types of validity, namely:

a) Content Validity

In a journal put forward by Puspitasari and Febrinita (2021), content validity is a form of validity that is often used in research. It assesses the extent to which a research instrument, such as a test or questionnaire, covers all the aspects it intends to measure. In practice, content validity is tested by asking an expert in the relevant field to assess how well the instrument covers the material it is intended to measure. In other words, if the expert confirms that the instrument takes into account all the important matters in the field of research, then the content validity of the instrument is considered high.

Content validity in this validity evaluates the extent to which the test can representatively measure teaching materials or behavioral changes. It reflects how effective the items are in testing the appropriate material, through logical and rational evaluation to ensure compatibility with the instructional objectives and the material to be tested.

b) Criterion Related of Validity

Criterion related of validity, also known as criterion-based validity, is a measure of validity established by comparing test scores to specific performance on an external measure. A theoretical relationship is expected to exist between this external measure and the variable being measured by the test. For example, an intelligence test may be correlated with academic grade point average. (Hendryani, 2017)

c) Contract validity

Hendriani & Suzanne (2013), said that construct validity indicates how well the test measures a particular aspect. Factors such as students' motivation, their responses, reading ability, competitive nature, and the psychological quality of the test can affect the results. Students who are less motivated, have poor test psychological conditions, or are less competitive may perform less optimally. In addition, students who know the answers but misunderstand the questions, perhaps due to cultural differences or reading difficulties, may answer incorrectly.

2) Criteria for Validity

Validity describes how well a measuring instrument can achieve its purpose accurately and consistently. When a test or measuring device provides results that are consistent with the purpose of the measurement, this indicates a high level of validity, which means that the results reflect the actual facts or conditions. If the calculation result is less than 0.3, the question is considered invalid, while if the result is 0.3 or more, the question is considered valid.

The implementation of validity theory is very important to help teachers understand and identify deviations in question formulation. Deviations occur when the questions created do not align with basic competencies, learning indicators, or expected cognitive levels. For example, if a teacher formulates tests that are too easy or only measure memorization, whereas the indicators direct towards understanding or analysis, then this is a form of deviation that reduces the validity of the questions. In addition, validity can also be tested statistically through the correlation between the scores of each item and the total score. Questions that have low validity (usually below 0.30) indicate that they do not provide significant contributions to the overall test results. Thus, the application of the concept of validity allows teachers to assess the quality of questions objectively and identify any deviations that may occur in the test formulation process.

Thus, the concept of validity allows teachers to objectively assess the quality of questions and to identify any deviations that may occur in the test preparation process. Understanding validity not only helps improve the quality of the instruments but also ensures that the test results truly reflect students' abilities fairly and accurately.

b. Reliability

1) Definition of Reliability

According to Brown & Abeywickrama (2004), in Language Assessment Participle, a reliability test is one that provides stable and trustworthy results. That is, if the same test is given to the same student or students with similar profiles at two different times, the results will be consistent or almost the same.

A reliability test is at the core of the validity of a measuring instrument. In this situation, "reliability" refers to the constancy of the results obtained from the instrument when used on the same subject or subjects with similar profiles at two different times. Thus, if a test is considered reliability, the results obtained tend to be fixed and consistent over time. This indicates that the test provides a reliable measure of what is being measured. In simple terms, the measuring instrument can be relied upon to provide consistent results, which form a solid basis for interpreting and making informed decisions based on the test results.

Reliability is the extent to which a test provides consistent results when used repeatedly under relatively similar conditions. Arikunto (2013) states that an instrument has high reliability if it produces stable or consistent results even when used repeatedly. Reliability is very important because without consistent results, the test cannot be trusted to accurately depict students' abilities. In the context of question preparation, reliability can be tested using formulas such as KR-20 or KR-21 for multiple-choice tests. If the reliability test results are low, there may be inconsistencies or discrepancies in the tests, such as differing levels of difficulty or irrelevant tests. Deviation can occur when teachers do not pay attention to the consistency among questions or do not conduct reliability tests, resulting in test results that do not stably reflect students' actual abilities.

2) Criteria of Reliability

The reliability coefficient, ranging from 0 to 1, indicates the strength of the connection. When the coefficient approaches 1, it's considered reliable, and vice versa. Typically, a reliability coefficient of 0.6 or higher is seen as the minimum standard. If the calculated result falls below 0.70, the question is deemed unreliable, while a result of 0.70 or higher signifies reliability.

c. Level of Difficulty

1) Definition of Level of Difficulty

According to Zainul et al (1997), the level of difficulty is a test item is essentially about how many people answer it correctly, typically represented by p. A higher p value means more people got it right, indicating that the item is easier. Conversely, a lower p value suggests that the item is more difficult because fewer people answered it correctly.

The difficulty level of a question is a way to find out how difficult or easy the question is. If most students can answer correctly, then the question is considered easy, but if most students cannot answer it correctly, then the question is considered difficult. The average score obtained by students on the question reflects its level of difficulty.

2) Criteria Level of Difficulty

There are several criteria used in difficulty levels:

a) Items with a difficulty level between 0.00 to 0.30: difficult items, meaning they may require a higher level of thinking or skill to answer correctly

b) Items with a difficulty level between 0.31 to 0.70: moderate items, meaning they require a fair amount of skill or knowledge to answer.

c) Items with a difficulty level between 0.71 to 1.00: easy items, meaning they can probably be answered with relative ease by most people.

The level of difficulty is the extent to which a question is easy or difficult for students. The level of difficulty is calculated based on the proportion of students who answer a tests correctly. According to Sudijono (2011), the ideal difficulty index is between 0.30 and 0.70, meaning the tests are neither too easy nor too difficult. Tests that are too easy (above 0.70) or too difficult (below 0.30) cannot adequately assess students' abilities. In practice, deviations often occur when teachers formulate tests without considering this level of difficulty, for example, creating tests that are too simple or too complex for students. By applying difficulty level analysis, teachers can identify which tests need improvement to achieve an appropriate balance in measuring students' abilities.

d. Discrimination Index

1) Definitions of Discrimination Index

The discrimination index of an item can be determined by looking at the item discrimination index value, which is a measure of the distinguishing power of an item. The item discrimination index, generally denoted by the letter D, has a range of values between 0 and 1.00. The analysis of the level of difficulty and discriminating power in the objective form and description form questions is carried out differently.

2) Criteria of Discrimination Index

The criterion of discrimination is that the discrimination index of an item describes how well the item separates students who are good and not good. If all clever students answer correctly and all non-proficient students answer incorrectly, then the discrimination index of the item is 1.00. Conversely, if all clever students answer incorrectly and all non-proficient students answer correctly, then the discrimination index value is -1.00. and symbolized d.

The discrimination index indicates the ability of the tests to differentiate between students with high and low abilities. Arikunto (2013) explains that tests with a good discrimination index will be answered correctly by high-ability students and answered incorrectly by low-ability students. The discrimination index is calculated based on the difference in proportions between the upper group and the lower group in answering a question correctly. The ideal value is above 0.40. If the discrimination index is low or even negative, the test is considered ineffective because it fails to distinguish the level of students' abilities. Deviations occur when teachers compose tests without conducting this analysis, resulting in many tests that did not provided useful information about the differences in students' abilities in the class.

e. Distractors

1) Definition of Distractors

According to Amalia & Widayati 2012), multiple-choice questions differ from description questions in that multiple-choice questions have several answer options already provided. Among these answer options, only one is correct, while the others are distractors. These incorrect answer options are called distractors, which aim to divert or confuse the examinee. A good question will have distractors that are evenly distributed in the selection of participants' answers, while a poor question will have distractors that are selected unevenly.

It can be concluded that the placement of distractors on each item aims to attract the interest of diverse test takers, as well as differentiate abilities among them. Effective distractors will trick less skilled test takers, while proficient test takers will avoid them. Therefore, the success of a distractor can be measured by how attractive it is to test takers.

2) Criteria of Distractors

Effective answer pattern analysis is done by evaluating test takers' preferences for each answer choice in each item. In order to determine the criteria for a good exception, it is important that the option is chosen by at least 5% of the test takers. For example, the results of a pilot test of a learning instrument on 100 test takers showed that an outlier is considered effective if it is selected by at least 5 test takers.

Distractors are alternative answer choices apart from the correct answer in multiple-choice questions, which serve to divert the attention of students who do not understand the material well. A good distractor should be chosen by some students who have less understanding, not by students who have mastered the material. According to Haladyna et al. (2002), distractors that are not chosen at all or chosen by students randomly indicate that the distractor is not functioning well. In practice, many deviations occur as teachers focus only on the correct answer, without paying attention to the quality and logic of other choices. Poor distractors can cause questions to be too easy or confusing. Therefore, analyzing the effectiveness of distractors is very important to ensure that all choices in the tests function optimally in measuring students' understanding.

C. Conceptual Framework

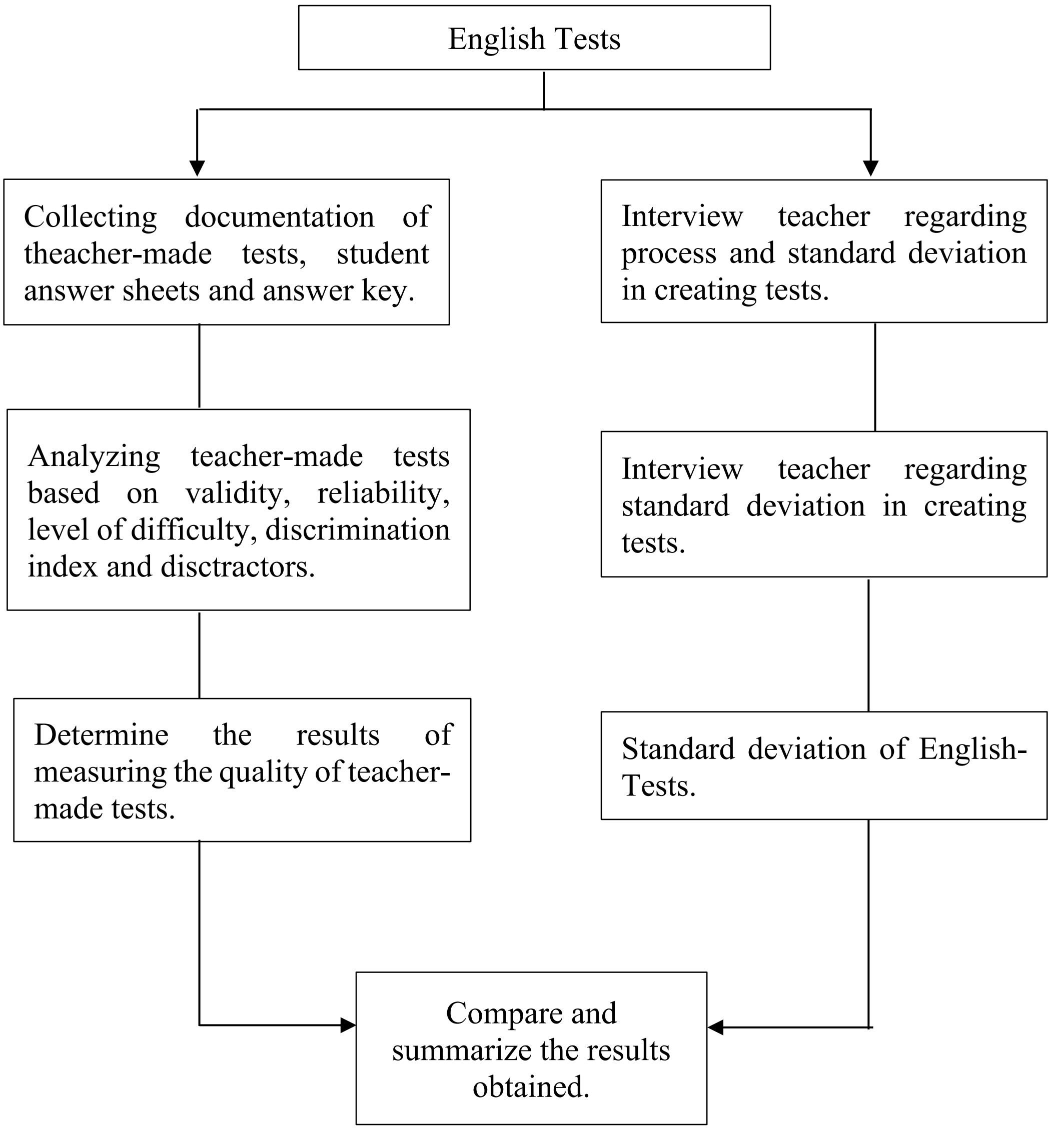
English Tests
Collecting documentation of theacher-made tests, student answer sheets and answer key.
Interview teacher regarding process and standard deviation in creating tests.
Analyzing teacher-made tests based on validity, reliability, level of difficulty, discrimination index and disctractors.
Interview teacher regarding standard deviation in creating tests.
Determine the results of measuring the quality of teacher-made tests.
Standard deviation of English-Tests.
Compare and summarize the results obtained.

Figure 2.1 Conceptual Framework

CHAPTER V
CONCLUSION AND SUGGESTION

A. Conclusion

Based on the results of research, an analysis of the English tests for second-grade students at SMP Muhammadiyah Mamuju and the teachers' deviation in constructing English tests. This research was conducted in class VIII.1 at SMP Muhammadiyah Mamuju, with 20 items of multiple-choice questions and interviews with one English teacher.

1. The quality of the multiple-choice English test is based on validity, reliability, level of difficulty, discrimination index, and distractor. The results of the analysis based on validity showed that only 11 questions (55%) were valid, and 9 questions (45%) were invalid. This shows that most of them are not able to measure students' competencies according to the learning objectives. The test reliability value of 0.51 indicates that the test is unreliable or inconsistent if used repeatedly. From the aspect of difficulty level, most of the questions were categorized as easy 12 questions (60%), moderate 7 questions (35%), and only a few were difficult 1 question (5%), indicating a lack of variety and sharpness of measurement. The discrimination index of the questions was not optimal, with only 9 questions (45%) classified as good, 7 questions (35%) satisfactory and 4 questions (20%) worst, so they were not able to distinguish effectively between high and low ability students. Meanwhile, the distractor in the questions showed varying results, with 12 questions (60%) very good, 5 questions (25%) good, and 3 questions (15%) quite good. Overall, these results indicate that the analyzed questions have not met the learning quality standards, and further revisions and training for teachers in preparing and developing good questions are needed.
2. English teachers face various obstacles in preparing quality test instruments, such as difficulties in designing effective distractors, determining the difficulty level of questions that are still subjective, and the absence of in-depth item analysis. In addition, fairness and consistency are not guaranteed due to the lack of technical procedures such as peer review. The questions created also did not cover all cognitive levels and did not consider students' abilities thoroughly due

to the lack of diagnostic data. These findings indicate the need for ongoing training and technical support for teachers.

3. Based on the results of descriptive statistical analysis, the standard deviation value of 2.8 indicates that the distribution of student scores is in the low category. This means that students' abilities in working on problems tend to be homogeneous and do not vary too much. Although the homogeneity of scores can reflect consistency, this condition also indicates that the questions given have not been able to distinguish students' ability levels optimally. A standard deviation that is too low can be a sign that the questions tend to be too easy or do not vary enough in difficulty. This is in line with the results of other analyses that show the low validity and discrimination index of the questions. Thus, this low standard deviation is one indicator that the quality of the questions needs to be improved in order to provide a more accurate and comprehensive picture of students' abilities. Thus, further training or assistance is needed in the preparation and analysis of questions to improve the quality of learning assessment.

Overall, while the test was valid in measuring the learning objectives, the low reliability and small standard deviation indicate the need for improvements in item construction to better explore the differences in student ability levels and produce a test that is more effective in assessing student achievement.

B. Suggestion

Based on the results of the analysis of multiple-choice tests and teacher interviews related to the creation of English tests conducted at SMP Muhammadiyah Mamuju class VIII 1, the suggestions that can be proposed are as follows:

1. For teachers: Teachers are advised to analyze tests more often before they are given to students. This analysis includes validity, reliability, difficulty level, discrimination index, and distractors.
2. For schools: Schools can provide training to teachers on good test development, including question writing techniques, selection of appropriate question types, and test analysis.

BIBLIOGRAPHY

Adom, D., Mensah, J. A., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. International Journal of Evaluation and Research in Education (IJERE), 9(1), 109–119. https://doi.org/10.11591/ijere.v9i1.20457

Anggriani, E. (2021). An analysis of multiple-choice items of English final semester test made by English teacher, Universitas Islam Negeri Alauddin Makassar. UIN Alauddin Makassar Repository.

Amalia, A. N., & Widayati, A. (2012). Jurnal Pendidikan Akuntansi Indonesia, 10(1).

Arikunto, S. (2013). Dasar-Dasar Evaluasi Pendidikan (Edisi Revisi). Jakarta: Bumi Aksara.

Arifin, Z. (2013). Evaluasi pembelajaran. Bandung: PT Remaja Rosdakarya.

Badan Pengembangan dan Pembinaan Bahasa. (n.d.). Kamus Besar Bahasa Indonesia. Diakses pada 11 Juni 2024. https://kbbi.web.id/simpang.

Bloom, B. S. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. New York: Longman, Green and Co.

Braun, H., Kanjee, A., Bettinger, E., & Kremer, M. (2006). Improving education through assessment, innovation, and evaluation. The American Academy of Arts and Sciences. http://hdl.handle.net/20.500.11910/6009

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

.

Darmawan, S., Rianti, Y., Yuliani, S., & Sumarni. (2022). A test-items analysis of English teacher-made test. Journal of English Education and Teaching (JEET), 6(4), 498-513

Djaali, & Muljono, P. (2008). Pengukuran dalam bidang pendidikan. Jakarta: Grasindo.

Gronlund, N. E. (2000). Assessment of student achievement (7th ed.). in Allyn & Bacon (Eds). Boston, MA.

Hakim, L., & Irhamsyah. (2020). The analysis teacher-made test for senior high school at State Senior High School 1 Kutacane. Journal Ilmiah DIDAKTIKA, 21(1), 10–20.

Hartati, N., & Yogi, H. P. S. (2003). Item analysis for a better quality test. ELIF Journal, 2(1), 59–70. https://jurnal.amj.ac.id/index.php/ELIF.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15(3), 309–334.

Hendriani, S., & Suzanne, N. (2013). Language testing (Cet. 1). Batusangkar: STAIN Batusangkar Press.

Hendryadi (2017). Validitas dan reliabilitas: Instrumen penelitian kuantitatif. Jurnal Riset Manajemen dan Bisnis (JRMB) Fakultas Ekonomi UNIAT, 2, 169–178.

Inanna, I., Rahmatullah, R., & Hasan, M. (2021). Evaluasi pembelajaran: Teori dan praktek. Makassar: Tahta Media Group.

Indrayani, M. S. D., Marhaeni, 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), 283–289.

Ismail, F., Astuti, M., & Sholikhah, H. A. (2020). Evaluasi pembelajaran berbasis riset. Palembang: Karya Sukses Mandiri.

Nitko, A. J., & Brookhart, S. M. (2014). Educational Assessment of Students (7th ed.). Boston: Pearson Education.

Marsevani, M. (2022). Item analysis of multiple-choice questions: An assessment of young learners. English Review: Journal of English Education, 10(1). https://journal.uniku.ac.id/index.php/ERJEE

Mulyadi, M. (2012). Konstruksi realitas dalam media massa. Jurnal Studi Komunikasi dan Media, 16(1). https://doi.org/10.31445/jskm.2012.160106

National Center for Education Statistics. (2019). Glossary of assessment terms. https://nces.ed.gov/programs/coe/glossary.

Nurgiyantoro, B. (2013). Penilaian dalam pengajaran bahasa berbasis kompetensi. Yogyakarta:Jurnal Faktor M, 4(1). https://doi.org/10.30762/factor_m.v4i1.3404.

Purnama, E. R. D., & Martubi. (2017). Analisis butir soal ujian akhir semester mata pelajaran pemeliharaan sistem kelistrikan otomotif dan mesin otomotif (PSKOMO) di SMK Taman Siswa Jetis. Jurnal Pendidikan Teknik Otomotif, Edisi XXI(1). https://journal.student.uny.ac.id/index.php/otomotif-s1/article/view/10184?utm.b.

Puspitasari, W. D., & Febrinita, F. (2021). Pengujian validisi isi (validitas isi) angket persepsi mahasiswa pembelajaran daring. Journal Focus Action of Research Mathematic (Factor M), 4(1), 77–90.

Qodir, A. (2017). Evaluasi Dan Penilaian Pembelajaran. (Ngalimun, Ed.). Yogyakarta: K-Media.

Rahman, A. A., & Nasryah, C. E. (2019). Evaluasi pembelajaran. Uwais Inspirasi Indonesia.

Ropii, M., & Fahrurrozi, M. (2017). Evaluasi hasil belajar (Syukrul Hamdi, Ed). Lombok: Universitas Hamzanwadi Press.

Sudijono, A. (2011). Pengantar Statistik Pendidikan. Jakarta: Rajawali Pers.

Sudjana. (2005). Metoda statistika (Edisi ke-6). Bandung: Tarsito.

Sugiyono. (2017). Metode penelitian kuantitatif, kualitatif, dan kombinasi (mixed methods). Bandung: Alfabeta.

Sugiyono. (2018). Statistik nonparametris untuk penelitian. Bandung: Alfabeta.

Tritschler, K. A. (Ed.). (2000). Barrow & McGee's practical measurement and assessment. http://ijere.iaescore.com.

Uno, H. B., & Koni, S. (2012). Asesmen pembelajaran. Jakarta: Bumi Aksara.

Wahyuningsih, E. T. (2015). Analisis butir soal tes objektif buatan guru ulangan semester ganjil mata pelajaran ekonomi kelas X di SMA Negeri 1 Mlati. Universitas Negeri Yogyakarta.

Widoyoko, E. P. (2014). Penilaian hasil pembelajaran di sekolah. Yogyakarta: Pustaka Pelajar.

Widyaningsih, N. W. N. (2012). The Analysis of sumative Test Made by Indonesian Language Teacher of Eleventh Grade of Science Class, Universitas Pendidikan Ganesha. Undiksha Repository.

Zainul, A., & Nasoetion, N. (1997). Penilaian hasil belajar. Pusat Antar Universitas, Direktorat Jenderal Pendidikan Tinggi, Departemen Pendidikan dan Kebudayaan.