Language Test Item Analysis Techniques

Teknik Analisis Item Tes Bahasa

Najmalia Fitra
State Islamic Institut of Parepare, Indonesia
Email: najmaliafitra0@gmail.com

Herdah Herdah
State Islamic Institut of Parepare, Indonesia
Email: herdah@iainpare.ac.id

Amira Ezzat Mahrous
Cairo University, Egypt
Email: amira.ezzatt@gmail.com

DOI: 10.14421/almahara.2025. 0111.09

Abstract

The quality of a learning outcome exam is heavily influenced by the quality of its items, hence it is vital to assess the test items to enhance its quality. Language learning is a complicated learning process that involves four skills: listening, speaking, writing, and reading. The challenge in this study is determining which tactics are used to assess language examinations and how to use them. The research methodology used is library research, which entails acquiring data from a variety of scientific journals for the study. The study's findings suggest that test items may be analyzed in two ways: qualitatively and quantitatively. In qualitative analysis, four variables are considered: material analysis, question creation analysis, cultural/language analysis, and test accuracy about student ability. In addition, quantitative research looks at the test's internal properties, such as test validity, reliability, difficulty level, test item discrimination power, and distractor function efficacy.

Keywords: *Analysis Techniques, Language, Test Items.*

ملخص

جودة اختبار مخرجات التعلم تتأثر بشكل كبير بجودة فقراته، لذلك من الضروري تقييم فقرات الاختبار لتحسين جودته. تعلم اللغة عملية معقدة تشمل أربع مهارات: الاستماع، والتحدث، والكتابة، والقراءة. التحدي في هذه الدراسة هو تحديد الاستراتيجيات المستخدمة في تقييم اختبارات اللغة وكيفية تطبيقها. المنهجية البحثية المتبعة هي البحث المكتبي، والذي يتضمن جمع البيانات من مصادر متنوعة من المجلات العلمية. تشير نتائج الدراسة إلى أنه يمكن تحليل فقرات الاختبار بطريقتين: تحليل نوعي وتحليل كمي. في التحليل النوعي، تؤخذ في الاعتبار أربعة متغيرات: تحليل المادة، وتحليل صياغة الأسئلة، والتحليل الثقافي/اللغوي، ودقة الاختبار بالنسبة لقدرات الطلاب. أما التحليل الكمي، فيركز على الخصائص

الداخلية للاختبار مثل الصدق، والثبات، ومستوى الصعوبة، وقوة تمييز فقرات الاختبار، وفعالية البدائل المشتتة.

الكلمات المفتاحية: تقنيات التحليل، اللغة، فقرات الاختبار.

Introduction

Test item analysis is an activity that must be carried out by teachers to improve the quality of questions that have been written. This activity is the process of collecting information from students' answers, summarizing it, and using it as a report for each evaluation. Because one of the functions of good questions is to help teachers find out which students have tried to prepare for the exam and which students have not. The purpose of item quality analysis, according to Aiken, is threefold: First, check each question before it is utilized to ensure quality questions; second, help improve questions by changing or removing ineffective questions; and third, identify students who grasp the topic and those who do not. As a result, the primary goal of teacher-led item quality analyses is to discover weaknesses in questions or learning.[1]

Various studies have emphasized the importance of qualitative and quantitative test item analysis to determine their validity and reliability. The ramifications of this test item analysis include improving the quality of test questions, resulting in more accurate and reliable evaluations. Tests play a distinct role in qualitative analysis and must be carefully evaluated. One critical aspect is the language and cultural fidelity of the questions. Furthermore, assessment of skills in language tests must cover various aspects such as listening, reading, writing, and speaking. Test item analysis must ensure that each skill is measured appropriately and proportionally. In quantitative test analysis, statistical data is used to evaluate the performance of each question item.[2]

Comprehensive test item analysis is performed by gathering, summarizing, and analyzing information from student responses to make conclusions regarding each exam. This item analysis contatestins validity tests, reliability tests, difficulty levels, discriminating

[1]Reni Lailina Hidayah et al., "Taḥlīl Bunūd Al Ikhtibār Al Nihā'i Limādati Al Lugah Al Arabiyah Fi Al Madrasah Al Ibtidā'iyah Al Islāmiyah Kebomlati Tūban," *Almahara* 7, no. 1 (2021): 1–26, https://doi.org/10.14421/almahara.2021.071-01.

[2]Muhammad Atanda Musa, Rara Mutiah, and Rahmani, "Analisis Butir Soal Bahasa Arab Di Mtsn Kota Parepare," *Sao Jurnal IAIN Parepare* 1, no. 2 (2022): 11–23.

power, and distractor functions.[3] Validity tests are used to measure the validity of a measuring instrument based on theoretical rationale and internal consistency. Reliability tests are used to appraise the reliability of questions, namely the ability to provide consistent results. Difficulty level tests are used to determine the level of difficulty of questions discriminating power tests are tests used to measure the ability of question items to distinguish between upper and lower groups, and distractor tests to determine whether the distractors of the questions function decently.[4]

The goal of test item analysis is to 1) give knowledge about the advantages and drawbacks of a question sheet, 2) provide thorough information about the specification of question items, 3) identify the problems found in the question items, 4) assess the collection of question banks, 5) as a reference for compiling question items. The results of a good test item analysis will provide an overview of the achievement of optimal student learning outcomes.[5]

Language item analysis test is critical to ensuring assessment quality and efficacy. Several studies have underlined the significance of analyzing items using qualitative and quantitative metrics to establish their validity and reliability. A critical component of item analysis is analyzing the content, structure, and linguistic appropriateness of the questions to evaluate their fit for educational purposes and capacity to measure student abilities reliably.

Quantitative item analysis evaluates each item's performance using statistical data. This method involves using difficulty indices, discriminating power, and distractor effectiveness to identify items that may be too easy, too difficult, or ineffective in differentiating between high and low-ability students. As a result, this analysis aids in identifying and improving problematic items. Expert judgment and the application of specified standards to assess the relevance and clarity of questions are common features of qualitative item analysis. This approach often includes analyzing the content to verify that all relevant curricular topics are covered, evaluating the question formulation to minimize ambiguity, and ensuring that the language used is appropriate for the students' level. This

[3]Iza Zainal Ambiya, Sopwan Mulyawan, and Hasan Saefuloh, "Analisis Soal Ujian Mata Pelajaran Bahasa Arab Kelas 12 Di Madrasah Aliyyah," *El-Ibtikar: Jurnal Pendidikan Bahasa Arab* 11, no. 1 (2022): 70–87.

[4]Ferdy Rakhmat Dianova and Najih Anwar, "Analisis Butir Uji Validitas , Reliabilitas , Tingkat Kesukaran , Dan Daya Pembeda Soal Sumatif Bahasa Arab SD Islam," *Jurnal Bahasa Daerah Indonesia* 1, no. 3 (2024): 1–13.

[5]Hendra Dani Saputra et al., "HASIL BELAJAR MAHASISWA : ANALISIS BUTIR SOAL TES," *Edukasi: Jurnal Pendidikan* 20, no. 1 (2022): 15–27.

involves taking cultural and linguistic factors into account while administering Arabic language assessments, which may impact understanding and performance.[6]

Implications of item analysis include increased test validity and reliability, which leads to a more accurate and reliable assessment of student skills. Furthermore, this research might help teachers identify areas where pupils may want further assistance or training. As a result, item analysis benefits not just evaluation but also the overall learning process. According to numerous literature research, evaluating questions must be understood and utilized in the educational process. A good test instrument will undoubtedly offer an objective image of the skills acquired by pupils during the learning process. and make a substantial contribution to raising educational standards, particularly in Arabic. As a result, the purpose of this study is to learn about and revisit the methodologies utilized in assessing language test items.

The research technique utilized in producing this article is library research, which is a way of gathering knowledge by analyzing and investigating theories from diverse readings linked to the topic. In this method, the first step is to collect literature sources that are relevant to the title of this research. These literature sources can all be books, scientific journals, articles, reports, and other reliable sources. Then, it is analyzed critically by considering its relevance, quality, and contribution to the research topic. The analysis is completed by reading, understanding, and grouping data based on themes, then connecting theories, arguments, or findings from various literatures. Relevant findings, theories, and patterns from the analyzed literature are then synthesized to form a comprehensive conceptual framework for analyzing language test items.

Results and Discussion

Language Test Item Analysis Technique

The language test item analysis technique is a procedure for analyzing test items. This test item analysis may be performed in two ways: qualitatively (qualitative control) and quantitatively (quantitative control). Qualitative analysis, also known as logical validity, is the analysis performed previous to the initial usage of a question to determine whether or not it works. Quantitative question analysis, also known as empirical validity, is the process

[6]Indah Rahmi Nur Fauziah, Syihabudin Syihabudin, and Asep Sopian, "Analisis Kualitas Tes Bahasa Arab Berbasis Higher Order Thinking Skill (Hots)," لسانـنا *(LISANUNA): Jurnal Ilmu Bahasa Arab Dan Pembelajarannya* 10, no. 1 (2020): 45, https://doi.org/10.22373/ls.v10i1.7805.

of testing question items on a population or sample to determine whether or not they function.

1. Qualitative Item Test Analysis

Qualitative analysis seeks to answer issues in terms of technical, content, and editorial elements. Technical analysis examines problems based on measuring principles and writing forms. Content analysis seeks to determine the feasibility of the knowledge sought, whereas editorial analysis seeks to determine the general structure and editorial consistency from one question to the next. The aspects studied qualitatively in language tests.[7]

a. Material aspect.

When analyzing the content, the applicability of the questions to the indicators and competencies to be measured must be evaluated. The material element of the Arabic language test must be analyzed in terms of its relevance to students' daily lives. Contextual and meaningful material will pique students' attention and boost their engagement. As a result, the question review should cover scenarios and themes that students will find relevant and fascinating. The material element of the questions is analyzed using the following criteria: 1) the questions are aligned with the indicators, 2) the content is aligned with the competencies, 3) the response options are homogenous and logical, and 4) there is only one answer key.

b. The Question Construction Aspect

The question creation component is the process of writing test items that already follow the standards of excellent question writing. Among the criteria analyzed in the question construction aspect are: 1) questions must be formulated briefly, clearly, and firmly; 2) questions do not provide clues to the correct answer; 3) questions do not contain negative questions; 4) pictures, lines, tables, diagrams, and the like must be clear and functional; 5) the length of the answer choices is relatively the same; 6) answer choices do not use "all answers are right/wrong" questions; and 7) question formulation.

c. Cultural dimensions of language.

Language analysis that seeks to investigate issues about the usage of excellent and proper language based on better spelling does not utilize the local regional language if the inquiries are intended for other areas or nationwide. The criteria for

[7] Musa, Mutiah, and Rahmani, "Analisis Butir Soal Bahasa Arab Di Mtsn Kota Parepare."

examining the language component of questions are: 1) using language that follows the rules, 2) utilizing communicative language, 3) not using local or regional language, and 4) response options that do not repeat the same terms.

In the context of language testing, the examination of test items includes certain subtleties that must be carefully studied. Language appropriateness is an important consideration. Languages, particularly Arabic, have a broad range of dialects and registers, thus it is critical that test items employ language that is acceptable for the setting of formal education and intelligible to all students. Linguistic mistakes or improper dialects can cause confusion and an incorrect assessment of student aptitude.

d. Adaptability of resources to students' abilities

Material analysis tries to assess the scientific substance of the questions while taking into account the degree of students' skills. By taking into account the content included in the core competencies and presented to students, the teacher has taught the material to the maximum extent possible in accordance with the expectations of the competencies that students must master, and the teacher is creative in presenting learning materials.

Expert review is frequently used in the qualitative analysis process to evaluate the questions' suitability, relevance, and clarity based on certain criteria. Before the questions are quantitatively examined, restricted trials are carried out if required to spot possible issues and make sure the questions are clear and don't lead to different interpretations. Before the test questions are utilized extensively, this qualitative analysis is crucial to ensuring their efficacy and quality so that the test results are reliable and legitimate.[8]

2. Quantitative Item Test Analysis

Quantitative analysis is carried out by testing the instrument that has been qualitatively examined on a group of students that have similar characteristics to the students who will be tested using the instrument. Quantitative question analysis focuses on analyzing the test's internal properties using empirical data. Internal

[8]Radu Bogdan Toma, Jairo Ortiz-Revilla, and Ileana M. Greca, "Development and Validation of a Multiple-Choice Test for Sustainability Competence in Primary School Using the GreenComp Framework," *International Journal of Educational Research Open* 7, no. April (2024): 100388, https://doi.org/10.1016/j.ijedro.2024.100388.

features measured quantitatively include validity, reliability, discriminatory power, level of difficulty, and distractor function efficacy.[9]

a. Validity Tes

Test validity must be evaluated to assess the test's ability to measure what should be tested. "valid" is defined as "accurate, correct, authentic, legitimate". According to Crocker and Algina (1986), validity is a metric that demonstrates whether a test or instrument is valid in measuring what needs to be assessed. It is also an assessment procedure that gathers empirical data to assist the proper interpretation and application of test results. Nana Sudjana defines validity as the assessment tool's correctness concerning the notion being assessed, ensuring that it truly examines what should be discussed.[10]

Based on the many definitions offered above, it is feasible to conclude that a test is genuine if it can measure what is intended to be tested. Meanwhile, the validity of a test item refers to the accuracy with which an item (a component of the test as a whole) assesses what should be evaluated. There are two types of validity: logical and rational.[11]

b. Logical Validity (rational)

Logical validity is a legitimate test based on reasoning outcomes carefully constructed in compliance with applicable ideas and requirements. Arikunto defines two categories of logical validity: content and construct validity.[12]

1) Material validity is determined by the analysis, tracing, and testing of the material presented in the learning outcome test. Substance validity is sometimes known as curriculum validity because it is concerned with the substance of the curriculum's materials. So the test is considered valid if it can represent the complete content.

---

[9]Khaerudin, "KUALITAS INSTRUMEN TES HASIL BELAJAR," *Jurnal Madaniyah* 2 (2015): 212–35.

[10]Fatimah Depi Susanty, "Analisis Validasi Soal Tes Hasil Belajar Pada Pelaksanaan Pembelajaran Bahasa Arab Di Pusat Pengembangan Bahasa (P3B) UIN SUSKA RIAU," *Kutubkhanah: Jurnal Penelitian Sosial Keagamaan* 19, no. 2 (2016): 124.

[11]Syaifudin, "Validitas Dan Reliabilitas Instrumen Penilaian Pada Mata Pelajaran Bahasa Arab," *Cross-Border: Jurnal Kajian Perbatasan Antarnegara, Diplomaasi Dan Hubungan Internasional* 3, no. 2 (2020): 106–18.

[12]Suharsimi Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, ed. Restu Damayanti, 2nd ed. (Jakarta: Bumi Aksara, 2013).

2) Construct validity refers to validity that is assessed based on its structure, framework, or design. A test has construct validity if the questions evaluate all aspects of thinking, such as knowledge, understanding, and application, in accordance with the competence standards, fundamental competencies, indicators, and learning objectives outlined in the curriculum.

c. Validity Empiric

Empirical validity is the assessment of measurement based on empirical analysis results, which is validity derived from or achieved by field observations. Validity may be determined empirically by making forecasts/predictions and comparing them.[13]

1) The validity of forecasts/predictions is a criterion indicating the capacity to foresee future events.
2) This comparative validity is also known as empirical validity, which means that the precision of measurement is based on field experience. A test is considered to have empirical validity if its results are consistent with experience.

Furthermore, there is a formula that may be used to calculate item validity, which is as follows: [14].

$$y_{pbi} = \frac{M_p - M_1}{S_t}\sqrt{\frac{p}{q}}$$

Description

y_{pbi} = Biserial correlation coefficient.

M_p = Is the average score of testees who correctly answered the question under consideration for validity.

M_1 = Average overall score.

S_t = Is the standard deviation of the overall percentage score.

p = The proportion of accurate answers

q = The proportion of students that replied wrongly (q= 1 – P)

$$p = \frac{number\ of\ students\ who\ answered\ correctly}{number\ of\ students}$$

[13]Arikunto.
[14]Arikunto.

In examining the validity of test items done with classical theory, the number of items and the number of respondents impact the outcomes of the analysis. Although this classical assessment is often employed in learning evaluation, it has limits in terms of the number of tests and samples. A minimum number of test items and a sufficient number of respondents are required for the results to be accurate and precise.[15] The lack of test items and respondents will result in low statistical test power, making it difficult to identify significant relationships or differences, the analysis's findings are often sensitive to outliers or extreme data. A small sample size can lead to Type I error, which is the mistake of rejecting a true null hypothesis. And Type II error, which is the failure to detect an effect that exists, and big samples are necessary for certain statistical tests (like ANOVA) to satisfy the homogeneity and normality requirements.

It is advised to double the number of objects to be used as research instruments in anticipation of the number that will be discarded. For example, if the study calls for 20 items, then 40 items, or twice as many, can be evaluated. Regarding the number of respondents, Crocker and Algina clarified that, despite the test having only 20 items, stability requires at least 200 respondents. In this instance, Crocker and Algina underlined the significance of sample size in ensuring reliable and accurate analysis outcomes, particularly in item analysis and model calibration.[16]

b. Tes Reliability

Reliability indicates constant values. A highly reliable test may be used to draw findings and make choices. A learning outcome test is considered excellent if it is trustworthy, which means that measurement findings obtained from the test on the same subject are consistent or steady. Anastasi (1976) describes reliability to the consistency of scores obtained by the same persons when reexamined the same test on different occasion, or with different sets of equivalent items or under other variable examining conditions.[17]

[15]Lai Kun Tong et al., "Validation of the Short Index of Job Satisfaction in Chinese Nurses: Classical Test Theory and Item Response Theory," *International Journal of Nursing Studies Advances* 8, no. August 2024 (2025): 100321, https://doi.org/10.1016/j.ijnsa.2025.100321.

[16]Idrus Alwi, "Kriteria Empirik Dalam Menentukan Ukuran Sampel Pada Pengujian Hipotesis Statistika Dan Analisis Butir," *Formatif: Jurnal Ilmiah Pendidikan MIPA* 2, no. 2 (2015).

[17]Zainal Arifin, *Evaluasi Pembelajaran*, vol. 8 (Bandung: Remaja Rosdakarya, 2017).

According to Karlinger (1986), stability, dependability, and predictability are the three criteria that can be used to quantify reliability. A test's stability indicates how consistently it measures the same symptoms. Predictability indicates a test's capacity to forecast the outcomes of follow-up symptom assessments, whereas reliability indicates a test's stability or degree of dependability. Among other things, increasing the quantity of test items can help a test become more reliable.

In simple terms, the usage of the test reliability approach is outlined in the table below:[18]

Table 1. Methods For Assessing Dependability

Reliable Methods	Procedure Implementation
Test-retest techniques (stability): product-moment correlation and intraclass correlation.	Present the identical test to the testee twice, at separate times, and then calculate the correlation.
Equivalent Parallel: product moment and intraclass correlation.	Present two identical tests to the same testee in a reasonably short time (e.g., two weeks), then compare the two results to determine reliability.
Split-half techniques (split in half); split-half and Sperman-brow equations.	Present the exam once, divide it in half, and then apply an equation to correlate the two parts.
Internal consistency; alpha coefficient, Kuder-Richardson (KR-20), Kuder-Richardson (KR-21).	Give one test and apply the equation. Give one test and apply the equation. Give the jam test, and utilize the equation.

Additionally, Gronlund (1985) provided thorough explanations of how four factors—test length, score distribution, difficulty level, and objectivity—can impact reliability.[19]

1) test's length is determined by how many questions it contains. There is a trend for the level of reliability to increase with test length. The reason behind this is that as the number of questions increases, more samples will be measured and a higher percentage of accurate responses will result, lowering the guessing factor.
2) Score distribution: When students stay in the same relative position in one test group from the next, a larger reliability coefficient is obtained; therefore, the greater the score distribution, the better the level of reliability.

[18]Khaerudin, "KUALITAS INSTRUMEN TES HASIL BELAJAR."
[19]Arifin, *Evaluasi Pembelajaran*.

3) Difficulty level: Both simple and complex questions typically yield low reliability scores in tests that employ the norm-referenced assessment approach. This occurs because there is a single, constrained score distribution for simple and difficult test results. Questions that generate a score distribution in the shape of a bell or normal curve are the best in terms of difficulty for raising the dependability coefficient.
4) Objectivity: Objectivity demonstrates that students' ability test results are identical. When students work on the same test, they will receive the same results if their skill levels are equal. High test procedure objectivity will yield reliable test results unaffected by the scoring process.

c. Difficulty Level

The difficulty level of the test items is computed by dividing the number of test takers who correctly answered the test items by the total number of participants. This level is represented by a test difficulty index (IKS) from 0.00 to 1.00. The higher the difficulty index derived from the computation results, the easier the question. This level of difficulty index is calculated for each test item number. Calculates the percentage of students who correctly answer the questions. Too high a score means the question is too simple, while too low a grade means the question is too challenging. A moderate level of difficulty is ideal for the questions in order to differentiate between students who have grasped the content and those who have not.[20] The level of the test items is typically tied to the aim of the exam. For example, if the test is intended for a semester exam, the questions are medium in difficulty; for selection, the questions are high in difficulty; and for diagnostic purposes, the questions are low/easy.

The formula for determining the level of difficulty for objective questions is.[21]

$$IKS = \frac{\mathrm{B}}{\mathrm{N}}$$

Description:

IKS = Question difficulty index.

[20]Michael J. Rudolph et al., "Best Practices Related to Examination Item Construction and Post-Hoc Review," *American Journal of Pharmaceutical Education* 83, no. 7 (2019): 1492–1503, https://doi.org/10.5688/ajpe7204.

[21]Abdul Munip, *Penilaian Pembelajaran Bahasa Arab* (Fakultas Ilmu Tarbiyah dan Keguruan UIN Sunan Kalijaga Yogyakarta, 2019).

B = Refers to the number of right responses among pupils.

N = Represents the total number of students who answered the question.

The Difficulty Level Index computation results can be interpreted as follows:

Table 2. Difficulty Level Index

Difficulty Level Index	Category
0,00 – 0,30	Difficult/hard
0,31 – 0,70	Medium
0,71 – 1,00	Easy

For example, 5 persons take a language test with 5 questions, and the score details are presented in the table:

Table 3. Test Score Details

Testee	The score for question item number				
	1	2	3	4	5
Arif	1	1	1	1	0
Bima	1	1	0	1	1
Wahyu	1	1	1	1	0
Rafli	1	0	1	1	1
Fadil	0	1	0	1	0
N =	4	4	3	5	2

Calculating the difficulty index of questions 1 and 5 based on the following details:

Item 1 $= \frac{4}{5} = 0.8$ The question falls within the easy category.

Item 5 $= \frac{2}{5} = 0.4$ The question falls into the medium category.

Meanwhile, to estimate the level of difficulty of essay questions, the following formula is used.[22]

$$Mean = \frac{\text{Total Student Scores one question}}{\text{number of students taking the exam}}$$

$$IKS = \frac{\text{Mean}}{\text{Maximum score for each question}}$$

For example, five participants take a descriptive language test and get the following score:

Table 4. Test Score Details

Testee	The score for question item number				
	1	2	3	4	5
Arif	8	5	4	7	5

[22]Munip.

Bima	5	8	7	5	3
Wahyu	7	9	8	6	4
Rafli	6	4	5	9	6
Fadil	4	5	3	5	7
Jumlah	30	31	27	32	25

The difficulty index of question number 1 is: $Mean = \frac{30}{5} = 6$; then $IKS = \frac{6}{10} = 0.6$; the question is in the medium category.

In this case, the proportion of test takers who correctly answer is less representative due to the limited number of respondents, which means that a small number of samples will produce an unstable evaluation of the level of difficulty. Determining whether the questions are easy, moderate, or difficult will also be tricky. The difficulty of measuring the test items is also significantly influenced by the quantity of samples. A small sample will also affect the reliability and discriminatory power analysis of the test items, which are directly related to their level of difficulty, reducing the credibility of the results. Additionally, extrapolating the results of the difficulty study to a broader population will be problematic, especially if the sample is not representative or homogeneous.[23]

d. The discriminatory power of test items

A test item's discriminating power refers to its capacity to discriminate between students in the intelligent group (upper group) and those in the lower group. According to Anas Sudijono, the discriminatory power of an item is the ability of a learning outcome test item to differentiate between testees with high ability (clever) and testees with low ability (stupid) so that most testees with high ability answer more questions correctly, as testees with low ability are primarily unable to answer the item correctly.[24] The discrimination index is determined by dividing the group into two parts: the upper group (test participants with high ability) and the lower group (test participants with poor ability), discrimination

[23]Muhammad Jundi, "Classical Test Theory in Analyzing Arabic Test Questions: A Descriptive Study on Item Analysis Research in Indonesia/ نظرية الاختبار الكلاسيكية في تحليل الأسئلة العربية: الدراسة الوصفية على بحوث تحليل بنود الأسلة في إندونيسيا," *ATHLA : Journal of Arabic Teaching, Linguistic and Literature* 4, no. 2 (2023): 85–105, https://doi.org/10.22515/athla.v4i2.7747.

[24]Bahrul Hayat, "Klasika : Program Analisis Item Dan Tes Dengan Pendekatan Klasik," *JP3I (Jurnal Pengukuran Psikoligi Dan Pendidikan Indonesia)* 10, no. 1 (2021): 1–11.

index is defined as the difference in the proportion of right responses in the upper or intelligent group versus the lower or less clever group.[25]

To calculate the discriminating power index of objective form questions, apply the following formula:[26]

$$IDB = \frac{\mathrm{BA} - \mathrm{BB}}{\frac{1}{2}\mathrm{N}}$$

Description

IDB = Discriminatory Power Index of Question Items

BA = Number of accurate responses in the top group

BB = The number of accurate responses in the bottom group

N = Number of students who took the test.

The discriminating power level index findings are interpreted using the following criteria:

Table 5. Discriminatory Power Index

Discriminatory Power Index	Classification	Category
- (negative)	No discrimination	No discrimination power
< 0,20	Poor	Weak discrimination power
0,20 – 0,39	Satisfactory	Sufficient discrimination power
0,40 – 0,69	Good	High discrimination power
0,70 – 1,00	Excellent	Excelllent diskrimination power

For example, six participants take a language test with ten questions, and the score details are as follows:

Table 6. Test Score Details

Testee	The score for question item number										Total score	Group
	1	2	3	4	5	6	7	8	9	10		
Arif	1	1	0	1	0	1	0	0	0	1	5	B
Bima	1	1	0	1	1	1	1	1	1	0	8	A
Wahyu	1	1	1	1	0	1	0	1	1	1	8	A
Rafli	1	0	1	1	0	0	0	0	1	1	5	B
Fadil	0	1	0	1	0	1	0	0	1	1	5	B
Syarif	1	1	0	1	0	1	1	1	1	1	8	A

[25]Umi Ma'rifah, Nyanuar Algiovan, and Cucu Sutarsyah, "An Item Analysis of English Test During Online Learning," *International Journal of Multicultural and Multireligious Understanding* 8, no. 12 (2021): 647–54.

[26]Munip, *Penilaian Pembelajaran Bahasa Arab*.

Next, compute the item power index for question number 2 as follows:

Item number 2 has an $IDB = \frac{3-2}{3} = 0.33$ indicating acceptable discriminating power.

The formula for assessing the discriminating power of essay questions is as follows:[27]

$$IDB = \frac{Upper\ group\ mean(MA) - Lower\ group\ mean\ (MB)}{\text{Maximum item score}}$$

For example, 5 participants take a descriptive language test and get the following score:

Table 7. Test Score Details

Testee	The score for question item number					Group
	1	2	3	4	5	
Arif	8	5	10	7	5	A
Bima	5	8	7	5	3	B
Wahyu	7	9	8	6	4	A
Rafli	6	4	7	9	6	A
Fadil	4	5	3	5	7	B

To calculate the item discrimination index, first compute the mean of the upper group and the mean of the lower group, especially for question number 3:
The item discrimination index for question 3 is:

$MA = \frac{10+8+7}{3} = 8.33$

$MB = \frac{8+3}{2} = 5.5$

$IDB = \frac{8.33-5.5}{10} =$ 0,28; hence, the item discrimination index falls into the adequate group.

Analyzing discriminatory power in small samples will present several challenges, such as: the small number of students makes the division of upper and lower groups less representative,[28] making the upper and lower groups very small and less statistically stable; the discriminatory power index value fluctuates greatly and does not reflect the actual ability; high data variation and small group sizes can make good questions have negative or zero discriminatory power, making it difficult to accurately determine the quality of the questions.

[27]Munip.

[28]Hidayah et al., "Taḥlīl Bunūd Al Ikhtibār Al Nihā'i Limādati Al Lugah Al Arabiyah Fi Al Madrasah Al Ibtidā'iyah Al Islāmiyah Kebomlati Tūban."

e. The Effectiveness of The Distractor Function

Distractor function analysis is performed on objective questions with multiple-choice models. Multiple-choice questions have alternative responses known as choices. These selections often vary from four to five, notably a, b, c, d, and e. Of these alternatives, one is the correct answer, known as the answer key, while the others are incorrect responses, known as distractors.

The primary goal of placing distractors on each item is to pique the curiosity of some of the numerous testees taking the language test, who believe that the distractor they select is the correct answer. The more testees are tricked, the better the distractor works. In other words, a distractor is considered to work well if it has a stimulating or enticing power that leads the testee (particularly those from the lower group) to select the distractor as the correct response.[29]

A distractor is considered to operate successfully if: 1) at least 5% of testees choose it, and 2) the lower group chooses it more frequently. Calculating the function of a distractor is quite similar to calculating the question difficulty index, which is the proportion of testees who select the choice divided by the total number of testees. For example, consider a multiple-choice question with four response alternatives (a, b, c, and d) and one answer key (a). A total of 25 students worked on the question, with the following response choices: Ten participants picked answer a, seven chose answer b, six chose answer c, and two chose response b. So:

1. Distractor B = 28%; has worked well.
2. Distractor C = 24%; has performed well.
3. Distractor D = 0.8%; not yet operational and must be altered.

According to Suharsimi Arikunto, the follow-up following the distractor analysis can be approached in three ways: 1) Accepted because it is good, which means that 5% of the test participants chose all of the distractors in the question; 2) Rejected because it is not good, which means that 0% of the test participants chose any of the distractors; and 3) Rewritten because it is not good, which means that the distractors did not perform their function correctly.[30]

Due to the small number of participants, the proportion of distractor selection is less steady and representative, which limits the usefulness of the

[29]Rudolph et al., "Best Practices Related to Examination Item Construction and Post-Hoc Review."

[30]Anida Rahmaini and Aditya Nur Taufiq, "Analisis Butir Soal Pendidikan Agama Islam Di Smk n 1 Sedayu Tahun Ajaran 2017/2018," *Jurnal MUDARRISUNA* 8, no. 1 (2018): 1–24.

distractor function in small samples. When at least 5% of test takers select a distractor, it is deemed effective. Because 5% is a relatively tiny sample size, the effectiveness of the distractor can be greatly impacted by just one or two voters. The results of the distractor effectiveness study in small samples should therefore be read cautiously, and they typically need to be complemented by qualitative analysis or retesting on larger samples.

2. Qualitative and Quantitative Analysis Item Test at Different Levels of Education

Qualitative and quantitative analyses of tests at various levels of education are used to get a full picture of the quality and usefulness of tests in assessing students' skills. The table below presents a qualitative and quantitative examination of tests at various levels of schooling.

Table 8. Qualitative and Quantitative Analysis Item Test

Education Level	Qualitative Analysis	Quantitative Analysis
Primary Education	• Examine questions for relevance to curriculum learning outcomes. • Examine question language for clarity and potential bias • Group questions by cognitive domains (C1-C3) based on student development levels	• Calculate question difficulty index and discriminatory power • Test validity and reliability. • Examine the distribution of simple, medium, and tough questions based on the desired percentage (30% easy, 40% medium, and 30% challenging).
Secondary Education (Junior and Senior High School)	• An experienced panel reviews the content and construction of the test. • Assessing the compatibility of indicators with the competencies being tested and matched to students' cognitive capacities (Bloom's C1-C6). • Examination of student and instructor answers to support qualitative data	• Statistical examination of difficulty levels, discriminating power discriminant, and reliability (Cronbach's Alpha) • Use statistical tools (e.g., ITEMAN, Anates) for item analysis. • Apply normality and hypothesis tests to assess learning results.
Higher education	• Review assessment rubrics and performance assessment instruments. • Conduct interviews and questionnaires to gather qualitative data from lecturers and students.	• Analyze quantitative data from exam results and performance assessments. • Used Item Response Theory (IRT) for item analysis • Tested assessment instruments for construct

	• Evaluate instrument suitability to learning objectives and academic standards.	validity and reliability.

This table indicates that the application of qualitative and quantitative analysis of tests is tailored to the characteristics of students and assessment objectives at each level of education, and employs suitable methodologies and techniques to develop valid, reliable, and effective test instruments. Each item has a significant impact on the test's quality (Sharma, 2000). Item analysis aids in identifying items that are either too easy or too tough, things that do not differentiate between students who have understood the content and those who have not, or questions that contain illogical distractions (Lange, Lehmann, & Mehrens, 1967). Such materials can be removed, altered, or corrected by teachers, who can also modify their methods of instruction to clear up any misunderstandings or confusion regarding the subject matter.[31]

3. Qualitative and Quantitatif Analysis for Interpretation of Language Test Result

Qualitative and quantitative analysis play complementary roles in interpreting language exam results, influencing how to perceive the quality of questions and students' overall ability.[32] The table below shows the qualitative and quantitative analysis used to analyze language test results:

Table 9. Qualitative and Quatitative Analysis for Interpretation

Aspect	Qualitative Analysis	Quantitative Analysis
Data Focus	Understanding the meaning, context, and quality of student responses; Examining mistakes, cognitive processes, and cultural influences on outcomes.	Statistically determining the scores, frequencies, and distribution of students' language exam results; numerically testing the reliability and validity of questions.
Interpretation Objectives	To develop a thorough grasp of how learners' mental processes, linguistic challenges, and socio-cultural circumstances affect language test results.	Use numerical and statistical data to form inferences about learners' linguistic ability.

[31]Suleiman Sa and Elizabeth Julius, "Item Analysis: A Veritable Tool for Effective Assessment in Teaching and Learning," *Journal of Education and Practice* 12, no. 21 (2021): 22–28, https://doi.org/10.7176/jep/12-21-04.

[32] Sa and Julius.

Results of Interpretation	A descriptive narrative that explains why participants replied the way they did, including mistake factors and cultural or language contexts.	Numbers, percentages, item indices, difficulty levels, and discriminatory power can be used to compare groups.
Limitation	Interpretation is subjective, making it difficult to generalize to a large population.	It does not investigate the causes for the replies, hence it is less comprehensive in comprehending pupils' issues.

By integrating these two methods, language test results may be interpreted more comprehensively, taking into account not just numbers and statistics but also the context and quality of the questions that impact the outcomes.

4. Limitation and future research direction

a. Limitation

1) Primary Data Limitations: Because this study is based on literature, it does not entail collecting primary data directly from respondents, so the conclusions of the analysis are theoretical and depend on the quality and relevance of the sources utilized
2) Variation in methods and study contexts; the literature reviewed uses various item analyses and different educational contexts, so that the synthesis results may be less specific to a particular level or type of test.
3) Limited focus on multiple-choice questions; most of the studies reviewed focus on the analysis of multiple-choice questions, so they do not fully cover analysis techniques for matching questions, short answers, or other forms of tests that are also important in learning evaluation.
4) The role of less-considered external elements, such as culture, language, and student characteristics, on test item analysis outcomes is not thoroughly examined.

b. Future Research Direction

1) Develop item analysis tools for descriptive questions, essays, matching, and examinations based on higher-order thinking skills (HOTS) for a more complete evaluation.
2) Integrating qualitative and quantitative analysis; integrating qualitative methodologies, such as language and cultural background study of questions,

with quantitative analysis to generate a more comprehensive and reliable picture of question quality. c. Use larger samples and settings; undertake empirical research with large and varied samples to assess the reliability of the item analysis approaches outlined in the literature and enhance the generalizability of the results.

3) Use of test item analysis technology and software, creating and testing item analysis apps or software that instructors and researchers may use to consistently and efficiently enhance question quality.
4) Investigating the impact of cultural and linguistic elements; more studies may look at how students' cultural and linguistic backgrounds affect the validity and reliability of test questions, particularly in multicultural and multilingual settings.

Conclusion

Test item analysis is the process of assessing learning result test items based on student test responses to establish the test's quality as a measure of student learning outcomes. Teachers utilize test item analysis to enhance the quality of their questions. This exercise entails obtaining, summarizing, and using information from students' replies to decide on individual assessments. The test item analysis methodology includes two methods: qualitative analysis and quantitative analysis. The purpose of qualitative analysis is to answer technical, content, and editorial problems. Four components of qualitative analysis are examined: material analysis, question creation analysis, cultural/language analysis, and test accuracy vs student skills. Furthermore, there is a quantitative analysis, which examines the test's internal properties such as test validity, test reliability, test difficulty level, test item discrimination power, and distractor function efficacy.The primary goal of item analysis in a teacher-created exam is to discover weaknesses in the test or learning. Based on this purpose, item analysis activities have numerous advantages, including determining whether the item's function is as expected, providing input to students about their abilities and as a basis for discussion materials in class, offering input to teachers about student difficulties, providing input on specific aspects for curriculum development, revising the material being assessed or measured, and improving question writing skills.

References

Alwi, Idrus. "Kriteria Empirik Dalam Menentukan Ukuran Sampel Pada Pengujian Hipotesis Statistika Dan Analisis Butir." *Formatif: Jurnal Ilmiah Pendidikan MIPA* 2, no. 2 (2015).

Ambiya, Iza Zainal, Sopwan Mulyawan, and Hasan Saefuloh. "Analisis Soal Ujian Mata Pelajaran Bahasa Arab Kelas 12 Di Madrasah Aliyyah." *El-Ibtikar: Jurnal Pendidikan Bahasa Arab* 11, no. 1 (2022): 70–87.

Arifin, Zainal. *Evaluasi Pembelajaran*. Vol. 8. Bandung: Remaja Rosdakarya, 2017.

Arikunto, Suharsimi. *Dasar-Dasar Evaluasi Pendidikan*. Edited by Restu Damayanti. 2nd ed. Jakarta: Bumi Aksara, 2013.

Dianova, Ferdy Rakhmat, and Najih Anwar. "Analisis Butir Uji Validitas , Reliabilitas , Tingkat Kesukaran , Dan Daya Pembeda Soal Sumatif Bahasa Arab SD Islam." *Jurnal Bahasa Daerah Indonesia* 1, no. 3 (2024): 1–13.

Hayat, Bahrul. "Klasika : Program Analisis Item Dan Tes Dengan Pendekatan Klasik." *JP3I (Jurnal Pengukuran Psikoligi Dan Pendidikan Indonesia)* 10, no. 1 (2021): 1–11.

Hidayah, Reni Lailina, Saadatud Darain, Ahmad Zainun Yusuf, and Nur Qomari. "Taḥlīl Bunūd Al Ikhtibār Al Nihā'i Limādati Al Lugah Al Arabiyah Fi Al Madrasah Al Ibtidā'iyah Al Islāmiyah Kebomlati Tūban." *Almahara* 7, no. 1 (2021): 1–26. https://doi.org/10.14421/almahara.2021.071-01.

Khaerudin. "KUALITAS INSTRUMEN TES HASIL BELAJAR." *Jurnal Madaniyah* 2 (2015): 212–35.

Ma'rifah, Umi, Nyanuar Algiovan, and Cucu Sutarsyah. "An Item Analysis of English Test During Online Learning." *International Journal of Multicultural and Multireligious Understanding* 8, no. 12 (2021): 647–54.

Muhammad Jundi. "Classical Test Theory in Analyzing Arabic Test Questions: A Descriptive Study on Item Analysis Research in Indonesia/ نظرية الاختبار الكلاسيكية في تحليل الأسئلة العربية: الدراسة الوصفية على بحوث تحليل بنود الأسلة في إندونيسيا." *ATHLA : Journal of Arabic Teaching, Linguistic and Literature* 4, no. 2 (2023): 85–105. https://doi.org/10.22515/athla.v4i2.7747.

Munip, Abdul. *Penilaian Pembelajaran Bahasa Arab*. Fakultas Ilmu Tarbiyah dan Keguruan UIN Sunan Kalijaga Yogyakarta, 2019.

Musa, Muhammad Atanda, Rara Mutiah, and Rahmani. "Analisis Butir Soal Bahasa Arab Di Mtsn Kota Parepare." *Sao Jurnal IAIN Parepare* 1, no. 2 (2022): 11–23.

Rahmaini, Anida, and Aditya Nur Taufiq. "Analisis Butir Soal Pendidikan Agama Islam Di Smk n 1 Sedayu Tahun Ajaran 2017/2018." *Jurnal MUDARRISUNA* 8, no. 1 (2018): 1–24.

Rahmi Nur Fauziah, Indah, Syihabudin Syihabudin, and Asep Sopian. "Analisis Kualitas Tes Bahasa Arab Berbasis Higher Order Thinking Skill (Hots)." *لسانـنـا (LISANUNA): Jurnal Ilmu Bahasa Arab Dan Pembelajarannya* 10, no. 1 (2020): 45. https://doi.org/10.22373/ls.v10i1.7805.

Rudolph, Michael J., Kimberly K. Daugherty, Mary Elizabeth Ray, Veronica P. Shuford, Lisa Lebovitz, and Margarita V. Divall. "Best Practices Related to Examination Item Construction and Post-Hoc Review." *American Journal of Pharmaceutical Education* 83, no. 7 (2019): 1492–1503. https://doi.org/10.5688/ajpe7204.

Sa, Suleiman, and Elizabeth Julius. "Item Analysis: A Veritable Tool for Effective Assessment in Teaching and Learning." *Journal of Education and Practice* 12, no. 21 (2021): 22–28. https://doi.org/10.7176/jep/12-21-04.

Saputra, Hendra Dani, Wawan Purwanto, Dedi Setiawan, Donny Fernandez, and Rido Putra. "HASIL BELAJAR MAHASISWA : ANALISIS BUTIR SOAL TES." *Edukasi: Jurnal Pendidikan* 20, no. 1 (2022): 15–27.

Susanty, Fatimah Depi. "Analisis Validasi Soal Tes Hasil Belajar Pada Pelaksanaan Pembelajaran Bahasa Arab Di Pusat Pengembangan Bahasa (P3B) UIN SUSKA RIAU." *Kutubkhanah: Jurnal Penelitian Sosial Keagamaan* 19, no. 2 (2016): 124.

Syaifudin. "Validitas Dan Reliabilitas Instrumen Penilaian Pada Mata Pelajaran Bahasa Arab." *Cross-Border: Jurnal Kajian Perbatasan Antarnegara, Diplomaasi Dan Hubungan Internasional* 3, no. 2 (2020): 106–18.

Toma, Radu Bogdan, Jairo Ortiz-Revilla, and Ileana M. Greca. "Development and Validation of a Multiple-Choice Test for Sustainability Competence in Primary School Using the GreenComp Framework." *International Journal of Educational Research Open* 7, no. April (2024): 100388. https://doi.org/10.1016/j.ijedro.2024.100388.

Tong, Lai Kun, Yue Yi Li, Yong Bing Liu, Mu Rui Zheng, Guang Lei Fu, and Mio Leng Au. "Validation of the Short Index of Job Satisfaction in Chinese Nurses: Classical Test Theory and Item Response Theory." *International Journal of Nursing Studies Advances* 8, no. August 2024 (2025): 100321. https://doi.org/10.1016/j.ijnsa.2025.100321.