



Castledown

 OPEN ACCESS

Language Education & Assessment

ISSN 2209-3591

<https://www.castledown.com/journals/lea/>

Language Education & Assessment, 8(1), 2237 (2025)

<https://doi.org/10.29140/lea.v8n1.2237>

Evaluating Research Reports on the Qualities of Tests of English Language Skills in Indonesian Schools: A Systematic Review



PATRISIUS ISTIARTO DJIWANDONO^a

DANIEL GINTING^b

^a*Universitas Ma Chung, Indonesia*
patrisius.istiarto@machung.ac.id

^b*Universitas Ma Chung, Indonesia*
daniel.ginting@machung.ac.id

Abstract

The teaching of English as a foreign language in Indonesia has a long history, and it is always important to ask whether the assessment of the students' language skills has been valid and reliable. A screening of many articles in several prominent databases reveal that a number of evaluation studies have been done by Indonesian scholars in the last 14 years. This paper reports a systematic review with the aim of critiquing those evaluation studies to see the soundness of their methods and their results. PRISMA framework was used to screen a large number of articles from the databases and to finally obtain 14 research papers published in various journals. The findings indicate that most of the studies were focused on the analysis of the items in multiple-choice tests, and on the content validity, reliability and construct validity of those tests. A further scrutiny revealed that many of these studies lacked methodological rigor, including the absence of expert judgment in content validation, limited application of psychometric frameworks such as Aiken's V formula, and insufficient procedures for construct validation.

While the measurement of the item difficulty, item discriminatory power, and distractors' efficiency were relatively adequate, the approaches to determining the content validity, construct validity, and reliability of the tests remained overly subjective and inconsistent. These findings highlight the need for improvements in language test research practices in Indonesia, including structured training for teachers in language assessment, the adoption of psychometric-based validation methods, and systematic involvement of expert judgment in test development processes

Keywords: language tests, content validity, construct validity, reliability, systematic review

Copyright: © 2025 Patrisius Istiarto Djwandono & Daniel Ginting. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within this paper.

Introduction

Evaluation and testing have always been an indispensable part in academic sphere. It helps the practitioners to answer a burning question: how far and how well have we achieved our objectives? In a school context, it is often translated into tests. Tests comprise a vital part in the education system because they enable the students to see how good they are academically and enable the teachers to know whether they have been successful in their teaching. Taken to a narrower context of English language tests, the tests help the students to see how well they have been able to use English and help the teachers to see whether they have been successful in teaching their students how to use English to communicate.

The teaching of English in Indonesia dates back to 1967 when a decree by the Minister of Education established that English is a foreign language that must be taught in Indonesian schools (Komaria, 1998). The curricula that spanned from 1946 to 2013 have always aimed at the mastery of English language for reading, listening, speaking and writing skills. English is taught to fulfill four functions, namely cognitive, instrumental, integrative, and cultural (Nababan, 1991), with the emphasis being on the receptive skills. The current context of language teaching in Indonesia, as Hamied and Musthafa (2019) argued, sees the need to strive even more for improvement in English language teaching because of the increasingly competitive world, the expectation to master advanced knowledge and modern skills, and the perennial problems that have been hampering the teaching effectiveness. These include insufficient resources in schools, limited access to up-to-date teaching materials, inadequate teacher training in language assessment, and a lack of understanding about effective testing principles among educators. These challenges extend to the realm of language testing, where the quality of assessments directly influences teaching practices and learning outcomes. It follows from here that more research needs to be done in order to see how language teachers assess their students' proficiency and determine their effectiveness of instruction.

The qualities of language tests in Indonesian schools have been a central focus in some studies conducted by Indonesian researchers. Some were done to see the validity and reliability of the English language tests, either at the school level or the national level, and some others tried to see the qualities of the test items. After more than ten years (2010–2024), what is needed is a systematic review of these studies in order to ascertain that they were carried out through rigorous procedures and turned out sound results. The findings of such systematic review would be a general profile of the qualities of the studies, which would point out areas to be improved for future studies along the same focus and inform decision makers about the credibility of those studies. Thus, the objective of our research was to evaluate some published studies by Indonesian scholars that were aimed to measure the validity, reliability and test item quality of existing English language tests in some Indonesian schools.

Meanwhile, it is to be noted that in the field of language testing, validity is the central feature that determines the quality and trustability of language tests. Establishing the validity of a test, in practice, has its own ramifications. The paper starts with a theoretical review of this vital element, and then presents several studies that point out to the present condition of language testing in Indonesian schools.

The paper starts with a theoretical framework about validity as the most overarching element in testing. It is followed by a literature review of a few studies on a similar topic. Following this is a description of the systematic review, which eventually takes the form of a series of evaluation of some research reports published in several journals that assessed the quality of language tests in Indonesian schools. The evaluation essentially points out the shortcomings of such studies as well as the areas of testing that Indonesian teachers need to improve significantly. The paper finally suggests future areas to be explored and emphasizes some critical points for improvement.

Theoretical Framework

The quality of a test is closely linked to test validity, which is arguably the must-have quality for any test. Therefore, all schools and teachers, in particular, have to ascertain that any kind of test they make for making decisions about the students are valid in the first place. Chapelle (2020), and Chapelle and Lee (2021) proposed several evidence that test developers have to present as proof of the validity of the tests they make. First is expert judgment on the test content. A panel of experts in the relevant skills being tested should be presented with the test items and they should comment on whether the test items have adequately measured the elements of the skills, or the content of a syllabus as far as a classroom test is concerned. The level of agreement between the experts should then be calculated using Aiken formula (Sujati et al., 2021), which is a formula for determining the content validity of a test. The coefficient generated by the formula will indicate whether the test items reflect the content being measured. However, in Indonesia, the scarcity of qualified experts in language assessment and limited institutional support for collaborative validation procedures often hinder this process, resulting in subjective and informal content validation practices (Sudaryanto et al., 2019; Ulwatunnisa et al., 2023; Emmiyati et al., 2023). Second is an analysis of test-takers when taking the test. This kind of introspective technique aims at identifying the test-takers' covert cognitive processes and match them with the construct that is being measured. In the Indonesian context, such techniques are rarely applied due to constraints such as large class sizes, limited time, and the absence of assessment literacy training for teachers, making it impractical to conduct in-depth cognitive interviews or think-aloud protocols (Rasyid et al., 2024; Suherman, 2022; Hamied & Musthafa, 2019). Third is to see if there is a strong positive correlation between the scores of a test and the scores on other tests that have been proven as valid. This method, while theoretically sound, is difficult to implement in Indonesian schools because comparable standardized tests are scarce, and schools tend to rely heavily on locally produced, teacher-made tests that lack psychometric validation (Lie, 2007; Ginting & Saukah, 2016; Rasyid et al., 2024).. Finally, there should also be a consideration of how the test influences the society of which the test-takers are part. Indonesian high-stakes exams, such as the National Examination, have been criticized for creating negative washback effects, encouraging rote learning and test-oriented instruction at the expense of communicative language use (Lie, 2007). In summary, although validation frameworks stress expert judgment, cognitive validation, correlational evidence, and social consequences, their application in Indonesia remains limited due to institutional constraints, a lack of qualified assessors, and an exam-driven culture.

Content validation and construct validation of tests require rigorous procedures and complex calculations of several important variables. Experts in construct validation like Genesee and Upshur (1996) and Chapelle (2021), in particular, admitted that this type of validation is complex because it is not easy to define or specify clearly the construct being measured. The construct being assessed should be defined in detail so as to show how an instrument for measuring the concept can be developed. According to Mackey and Gass (2022), attempts to arrive at a construct validation should find other tests which supposedly measure some factors or abilities related to the construct in question. In practice, this demands considerable methodological expertise and institutional commitment, both of which remain limited in the Indonesian educational system (Lie, 2007; Hamied & Musthafa, 2019).

By contrast, fields like medical education or psychological testing routinely employ factor analysis and correlation studies as part of their construct validation processes (Downing, 2003; Anastasi & Urbina, 1997). For example, in medical education, standardized clinical examinations undergo rigorous multi-phase validation involving expert panels, pilot testing, and statistical analysis to confirm the alignment of test content with defined competencies (Downing, 2003). Such comprehensive validation efforts are rare in Indonesian language assessment research, where studies often rely on the researcher's individual judgment without involving psychometric experts or advanced statistical tools (Rasyid et al., 2024;

Ginting & Saukah, 2016). These contextual limitations—including a shortage of trained personnel, restricted access to psychometric tools, large classroom sizes, and a pervasive culture of exam-centric teaching—continue to obstruct the application of internationally recommended validation practices in Indonesia (Lie, 2007; Rasyid et al., 2024). In short, despite the complexity of construct validation and its routine application in other fields, Indonesian language assessment research continues to face systemic barriers that limit the adoption of rigorous validation practices.

This ramification and its concomitant expertise is what Indonesian researchers interested in language testing studies have yet to arm themselves in order to generate a more solid and convincing evidence of validity of Indonesian school tests. In the following section, several validation studies conducted by Indonesian scholars are presented and critiqued in the light of the aforementioned principles. The theory of validity discussed above is used as a standard against which every reported research on test quality is reviewed. More specifically, efforts are made to see whether the studies have applied the principles of content validation and construct validation when examining the school tests and the English tests in the national examination.

Review of Literature

This section discusses studies that have been conducted in the past on the same topic, that is, the quality of language tests in Indonesia. It turns out that such research is very scarce, which indicates that a systematic review of language tests in Indonesia has not been in the best interest of Indonesian scholars.

A review of Indonesian teachers' knowledge about language testing was conducted by Nurdiana (2020). The study reviewed a number of past research that was done to identify the teachers' level of mastery of language testing principles. It was found that most teachers did not have sufficient mastery of language testing principles. While Nurdiana's study investigated the teachers' knowledge about language assessment, our present paper is focused on the evaluation of language tests made by teachers. The study by Nurdiana also used secondary or library research without a detailed explanation as to the criteria for selection and the selection procedures, while our research was done using a relatively more rigorous procedure of systematic literature review.

Another study along a similar line was by Suherman (2022) who studied 16 high school teachers' language assessment awareness by using questionnaires and interview. Although most of them were aware of the importance of validity and reliability and had applied the principles in their testing practice, they admitted that their knowledge and skills for creating good and valid tests were still not adequate. Our study would gain from Suherman's research an insight into why Indonesian school teachers seem to grapple with the language testing method.

Another study by Puspawati et al. (2024) reviewed 28 studies on the development of LAL (Language Assessment Literacy) of language teachers from many different countries. They found that most of the training programs equipped the teachers with theoretical knowledge of language testing and practical application of the theories to test construction with various formats. Unfortunately, of those 28 studies, none came from Indonesia, indicating a lack of LAL training for Indonesian teachers. Our study would present information on the overall skill of Indonesian teachers in constructing credible tests as measured by some research that we reviewed.

Another Systematic Literature Review by (Mariappan & Osman, 2023) concluded from 19 journal articles that there was growing interest in authentic assessment, and that teachers should have to be trained to conduct proper authentic assessment. While offering a result from a systematic review, this study did not delve into the quality of language assessment. Our study would complement the result by offering a study using the same method which revealed the quality of language tests.

One study using Systematic Literature Review by (Susyla & Jaya, 2023) addressed types of digital tools for language assessment. After presenting numerous digital tools for testing, the writers also underscored the importance of teacher training on using these tools. Nevertheless, this study did not describe previous investigation that focused on the quality of language tests in Indonesia. Our study, therefore, would present a profile of studies that focused on that issue.

In summary, the existing literature reveals a persistent gap between Indonesian teachers' awareness of language assessment principles and their practical ability to design valid and reliable tests. These earlier studies lacked methodological rigor in either selection criteria or validation procedures. This situation underscores the need for comprehensive systematic reviews, like the present study, which go beyond assessing teachers' knowledge to critically evaluate the validity and reliability of the language tests themselves—thereby addressing a crucial yet overlooked aspect of language assessment research in Indonesia.

Method

Research Design

This study applied systematic review research design and used PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework to obtain journal articles that report the evaluation of English language tests administered at secondary school level in Indonesia. A systematic review was deemed appropriate for the objective because it has the potential to reveal how much effort has been expended to study language tests in Indonesian schools and how well such studies have been conducted. The following Figure 1 of PRISMA displays the steps that were taken to conduct the systematic review:



Figure 1 PRISMA flow diagram of article selection process.

In the first phase, the scope of the search was determined. The recency of the publication, the language used, and the databases were the main consideration. It was decided to search only articles published between 2010 and 2024 that used English and were published in reputable databases, namely, SAGE Journals, ERIC, JSTOR, ProQuest, Taylor & Francis Online, and Google Scholar. A special caution was taken to avoid predatory journals. The writers decided to cover the 14-year period because significant changes in the curriculum that also impacted English language teaching took place during that period. In 2014, a newly implemented curriculum put emphasis on communicative competence and English authentic texts. Between 2015 and 2019, schools were given more authority to design English language tests in the National Examination. Throughout this period, receptive skills that included reading comprehension and listening comprehension have always remained the central focus of English language tests in schools.

The second phase was determining the search criteria. It was decided to use the search terms of “language tests”, “validity”, “reliability” and “Indonesia”. Once these search terms were applied to the databases, an initial screening based on the titles and abstracts followed immediately. Articles that reported language testing in languages other than English, or any testing endeavor in a field other than English language were removed. To efficiently organize and manage the identified articles, a reference management tool like Zotero was used. This allowed for easy sorting and tagging of relevant articles. As each article was imported into Zotero, it was categorized by database, and metadata (such as author, publication year, and journal) was captured. This ensured that only articles meeting the criteria were retained for further analysis.

The next phase was the assessment of eligibility of the remaining 130 articles. The articles were read more meticulously to determine their eligibility for inclusion. A few articles were filtered out at this stage due to their obviously weak methods. What remained from this stage were the 14 articles that met all of the criteria and the filtering processes. These articles were then scrutinized to find the details of methods they used for evaluating some tests in Indonesian schools.

To maintain the objectivity of the selection, the two reviewers conducted the selection independently based on criteria for inclusion, namely that the articles report an evaluation of English language tests in Indonesia, and that the articles report studies with no glaring methodological weakness. As for the exclusion criteria, studies were excluded if they met any of the following conditions: the study focused on languages other than English, such as Indonesian language proficiency tests. The study assessed English language tests outside the secondary school setting, including university-level assessments, corporate language training, or informal education. The study did not assess the validity, reliability, or effectiveness of the language tests. The study lacked a rigorous research design, clear data collection methods, or sound analysis. The study involved language tests that were not administered in Indonesia. The study was published in predatory journals or appeared as a duplicate entry from different databases. Disagreements between the reviewers were resolved by discussion and the reviewers subsequently agreed on which studies were to be included in the final version. The procedure for inclusion and exclusion stated above was followed exactly and recorded, and thus there is a clear record concerning the decisions taken.

Tables 1 and 2 below summarize the selection process:

Table 1 *Databases, key words, and number of records processed*

Database	Key words for searching	Number of records identified	Number of records removed
JSTOR	"language tests", "Indonesia", "validity"	13,499	13,421
Google Scholar	"language tests", "Indonesian", "school", "validity", "reliability"	17,300	17,288
ERIC	"language tests", "Indonesia", "validity". Filter was set to "English", and location was set to "Indonesia".	20	0
SAGE Journals	"language tests", "Indonesia", "validity"	136	0
Mendeley	"language tests", "Indonesia", "validity"	1435	1433

Table 2 *Number of records processed and the reason for exclusion*

Database	Number of records screened	Number of records assessed	Number of records used	Reasons for exclusion
JSTOR	78	78	0	No article discusses language testing in Indonesia.
Google Scholar	12	12	11	The test was for Indonesian language.
ERIC	12	9	1	The test being evaluated did not assess language skills.
SAGE Journals	136	30	1	The tests being evaluated were not used in Indonesia, or did not test language skills.
Mendeley	2	1	1	The abstract has clearly shown that the methodology was very weak.

Two other databases, namely ProQuest and Taylor & Francis Online, were actually included in the initial search but even with the key words the search did not turn out any result. Hence, they are not shown in the tables above. It should be acknowledged that this review was limited to a selection of academic databases. While these were chosen for their accessibility and reputation, excluding other international, regional, and grey literature sources may have narrowed the scope of findings. Future studies are encouraged to broaden the range of databases and sources to capture a more comprehensive view of English language testing practices in Indonesian secondary schools.

Results

Popularity of Multiple-choice Format and the Washback Effect

A number of researchers have made efforts to chart the condition of language skill assessment in Indonesian schools. This section describes and critiques some studies about the quality of language tests that teachers have been using to assess their students' language skills. It will be apparent from the

studies that in school testing, which are created by the language teachers, multiple-choice format is hugely popular among schoolteachers. Notwithstanding the difficulty of constructing multiple choice tests, many teachers seem to favor this format. Presumably, it is the ease of scoring the multiple-choice tests which contribute to its popularity among the teachers. A test whose answers can be scored simply by matching each of the answer with the answer key probably appeals highly to many teachers who are busy teaching several classes and doing the administrative work on a daily basis. However, this general tendency to incline toward multiple choice test is not without risk. Conland, (1986, as cited in Hollingworth & Drake, 2012, p. 5) stated that

“no multiple choice question can be used to discover how well students can express their own ideas in their own words, how well they can marshal evidence to support their arguments, or how well they can adjust to the need to communicate for a particular purpose and to a particular audience.”

In practical terms, as Allanson and Notar (2019) stated, multiple-choice tests also have some disadvantages. First, a multiple-choice test is not easy to construct. Many teachers with their busy schedule and load of administrative work could not spend long hours constructing reliable and valid multiple-choice tests with good discriminatory power and appropriate difficulty levels. As the two studies above by Conland (1986, as cited in Hollingworth & Drake) and Allanson and Notar (2019) suggest, teachers have yet to pay closer attention to the quality of their multiple-choice tests so as not to disadvantage the students whose grades are dependent on the scores they earn in the tests. Second, a multiple-choice test cannot measure learners' creativity and articulation of ideas, two aspects of skills that are prominent in productive language skills. Third, a multiple-choice test is likely to encourage test-takers' guessing and, in an era where almost everything is digitalized, it is prone to students' cheating because the students can cooperate with their classmates through their gadgets or find an assistance from prohibited sources like artificial intelligence. Unfortunately, as Allanson and Notar (2019) also noted, many teachers seem to push with multiple-choice tests, resulting in low-quality tests being used as the instrument in high-stake assessment.

The issue of washback effect is equally important in a review of nationwide tests. It essentially believes that the skills which are tested, and the test formats have an important bearing on the teaching learning activities that precede the tests. In other words, the tests influence the teaching. Thus, the emphasis on reading skills in the national examination, which is called *Ujian Nasional* (National Examination) and is usually developed by the Indonesian ministry of education, and the multiple-choice format in which the test is given will likely encourage teachers and students alike to spend great efforts on reading lessons and working out multiple choice tests.

It should be no surprising to see that they put in more effort in reading and listening than in speaking and writing skills. An investigation into the washback effect was done by Saukah and Cahyono (2015). They studied six teachers from three different high schools with different qualities to identify how the teachers responded to the change of status of the national examination of English language. The policy back then was to reduce the weight of the English language sub-test in the National Examination for determining high school students' graduation. So, what was once a high-stake test that determined students' graduation is now changed to low-stake test. Even with that kind of change, the results showed that any kind of examination held nationwide tended to influence teachers' teaching strategies in schools with limited resources. These six teachers put a lot of test practice, or 'teaching to the test', in their teaching to make sure their students were accustomed to the test formats and materials in the national examination. In short, the format of English National Examination has caused a more or less undesirable washback effect on teachers' assessment at the school level.

Studies on Formative and Summative Assessment

The first point of concern is the quality of formative and summative assessment by high school teachers. The tests under review below, as their names indicated, were designed to form the students' abilities, and to measure their overall mastery of the entire lesson materials they have learned. The tests were all given in multiple-choice format. Table 3 below sums up the reviewed studies, followed by a more substantive discussion of each:

Table 3 *Studies on formative and summative assessment*

Writers	Objective	Object of analysis	Comments
Semiun & Luruk (2020)	Determine the content validity, the reliability, item difficulty, item discriminatory power, and distractors' efficiency.	A 50-item summative test in multiple-choice format.	No experts' judgment was involved, and the decision about content validity was too general and subjective.
Hilaldy (2021)	Describe the content validity of English summative test.	A summative test for the academic year 2018-2019.	No experts' judgment was involved, resulting in overly subjective opinions.
Andriyani & Gustine (2022)	Describe the English formative tests.	One English teacher and 34 students.	No experts' judgment or rigorous analysis was used.
Rohmah (2019)	Determine the content validity, the reliability, item difficulty level, discriminatory power, and distractors' efficiency.	Twenty answer sheets from a multiple-choice test made by a classroom teacher.	Content validation did not involve experts' judgment; statistical computation was used to achieve the other objectives.
Sugianto (2020)	Determine the item difficulty level, item discriminatory power, and distractors' efficiency.	A set of 36 answer sheets of a 50-item summative test.	Computation by statistical analysis turned out sound results.
Maharani & Putro (2020)	Determine the item difficulty level, item discriminatory power, and distractors' efficiency.	As many as 151 answer sheets from 74 schools.	A software-assisted computation turned out convincing results.
Karim (2021)	Determine the item difficulty level, and item discriminatory power.	A set of 35 answers sheets to a 50-item summative test.	A computation using available statistical analyses turned out good results.

Andriyani and Gustine (2022) carried out a study on the formative tests for junior high school students at one particular school in West Java. The respondents were one English teacher and 33 students that the teacher had been teaching. Document study, observation and interviews were used to gather the data. On the basis of the average scores of the students, the researchers concluded that the "English formative test results are in the high category, this happens because the formative tests used are very reliable and valid" (Andriyani & Gustine, p. 308). The study could have turned out a sounder result if steps for ensuring the degree of validity and reliability, which could be quite stringent and rigorous, were presented after that above claim. In the future, studies on the validity of teacher-made tests still need to be conducted with a more rigorous method.

Another study in the same vein was done by Semiun and Luruk (2020). They examined the content validity of a multiple-choice summative test at one local school in Kupang; in addition, they also analyzed its reliability, difficulty level, discriminatory power, and distractors' effectiveness. The content

validity was measured by matching the test items with the objectives stated in the curriculum, while the rest of the test elements were analyzed by K-R 20 and item analysis. They found high content validity of the test, and that the test contained many items which were too easy but with satisfactory discriminatory power and distractors' effectiveness. This research, compared with the previous one, gave a relatively more credible picture of testing practice in schools albeit still centered on multiple-choice format. Upon closer scrutiny, however, it was clear that the approach to the content validation did not involve experts' judgment, and that the stringent criteria mentioned by the researchers for deciding on the validity were not fully applied. The resulting final decisions about the content validity, therefore, seemed coarse and too subjective.

Another study similar to this was done by Hilaldy (2021), which claimed that the test under study had appropriate content validity after the writer singlehandedly determined the match between the test items and the syllabus content. With only a few sentences claiming that the validity was satisfactory without enough empirical and statistical evidence, the result of his study was highly dubious. This lack of rigorous validation raises concerns about the credibility of the conclusions drawn from the test results. The absence of expert involvement in assessing content validity not only diminishes the reliability of the study's findings but also sets a poor example for future classroom-based assessment research. It implies that future studies must incorporate systematic and expert-driven validation processes to enhance the quality and trustworthiness of English test development at the secondary school level.

Several studies investigated the quality of teacher made-tests with the focus on the item difficulty level and item discriminatory power of multiple-choice tests. A recent study in this category was conducted by Rohmah (2019). The researcher analyzed 20 answer sheets from a private school in West Java and submitted them to analyses of its content validity, reliability, difficulty level, discriminatory power, and effectiveness of the distractors. The content validity was practically the result of the researcher's individual judgement and could have been made more objective by involving experts' judgment. The analyses of the reliability and item quality were rather straightforward with Anates version 4 as a means for statistical computation which turned out credible findings.

While the researchers claimed that the content validity and reliability were reasonably satisfactory, the item difficulty level was not good, with most of them being too easy or too difficult. Also, the test had weak discriminatory power and less effective distractors. The author underlined the importance of training for teachers in test development so that they were able to create valid and reliable tests.. However, it should be stressed that the study was focused on the quality of a multiple-choice test rather than a more communicative and authentic language test such as speaking or writing. Following the researcher's suggestion means having the teachers spend time and effort in developing multiple-choice format, a format which would hardly promote authentic language-based interactions between the test-takers.

Maharani and Putro (2020) conducted another item analysis involving a larger number of test sample. They analyzed 151 answers sheets from junior high school students from 74 schools in Ponorogo, East Java. They found that too many of the test items were of moderate and low difficulty. Most of the items had a relatively good discriminatory power, and 80% of the distractors were effective. Their analytical method was assisted by a software and therefore was quite effective in measuring the item qualities. Yet again, the study was done on a multiple-choice format, a widely popular format which somehow does not reflect the communicative nature of using the English language.

Sugianto (2020) conducted a study on the quality of the multiple-choice test items made by some English teachers in one of the top schools in Palangka Raya. The items were assessed for their difficulty, effectiveness of the distractors, and discriminatory power. He found that most of the items performed poorly on those parameters. This means that the items were not at the right level of difficulty for the

students. Many were too easy or too difficult, and both of these weaknesses rendered the test ineffective in measuring the target language skill. By poor distractors, it means that the options other than the right options in the items were too obviously incorrect in the students' eyes, prompting them to choose the right option without much thinking. Finally, poor discriminatory power indicates that the test items were unable to separate the high-achieving students from the low-achieving ones. A test which does not differentiate the outstanding students from the poor ones is not effective because the teachers will not be able to tell which students have mastered the lessons, and which ones have difficulties in mastering the lessons. In other words, teachers will not be able to determine if their teaching has been successful or not.

Still in the area of item analysis, Karim et al. (2021) carried out an analysis of reading comprehension test items made by a teacher for 35 students at a junior high school in West Nusa Tenggara province. They found that 32% of the items were poor in their difficulty level and discriminating power, and 24% were of mediocre quality. In this study, existing statistical analyses for computing the item difficulty level and discriminatory power were properly used. The researchers then underscored the importance of item analysis skill for high school teachers. It is believed that such skill will sensitize them to the quality of their teacher-made tests.

In short, some studies have set out to critically assess the quality of teacher-made classroom tests, all of which happened to be in multiple-choice format. The methodology used in the studies discussed above was reasonably sound. Their item analysis more or less pointed out that classroom tests have yet to be better in terms of content validity, discriminatory power and difficulty level. It can be concluded that many teachers of English in Indonesian schools still lack the ability to construct good multiple-choice tests. They would benefit from learning the principles of creating multiple-choice tests, which include making plausible distractors, measuring the depth of the test-takers' comprehension, and ensuring the reliability of the tests and the validity of the interpretation of the results (Haladyna, 2022). Thus, what needs to be done in the future is twofold: first, more intensive in-service training for language teachers need to equip them with sound knowledge of multiple-choice tests; second, studies that probe into the quality of teacher-made tests other than multiple-choice format also need to be conducted.

Studies on the Validity of the National Examination Test

Effendi and Suyudi (2016) noted that English national examination has been in place in Indonesian education system since the 1950s. The examination commonly used multiple-choice format for assessing the language proficiency of Indonesian high school students. With regard to this popular format in the national examination, this section critiques some research articles on the validity of the National Examination of English language for Indonesian high school students. Table 4 below summarizes the studies, followed by detailed discussion of each:

Table 4 *Studies on the validity of the national examination*

Writers	Objective	Object of analysis	Comments
Widyaningrum & Prabandari (2013)	Determine the content validity of English test in the National Examination.	The test items in the National Examination.	No description of the method was presented.
Rukmini & Jumaroh (2015)	Determine the content validity of test items in the National Examination.	Lesson objectives for Grade 10 to Grade 12, and the test items.	Findings were based on mere matching of the items against the prescribed skills in the curriculum; No deep analysis of the content was done.

(Continued)

Writers	Objective	Object of analysis	Comments
Geribka (2016)	Determine the content validity of the reading test section in the National Examination	The reading test section and the test specification.	Absence of expert judgment renders the result less credible, making its validity dubious.
Sabrina (2016)	Determine the quality of reading texts and comprehension questions in the National Examination; determine the item difficulty level and distractor effectiveness.	The reading texts and comprehension questions in the National Examination.	Lack of clear and rigorous method.
Jayanti, Husna, & Hidayat (2019)	Determine the content and construct validity of the National Examination.	Test items and the syllabus.	Lack of rigor in its method.

A study of content validity of the reading tests for senior high school students on the national examination was done by Widyaningrum and Prabandari (2013). The reading test on the national examination was matched against the Competence Standard and Basic Competence in the curriculum. The authors claimed that although the reading test items were entirely valid when matched against the Competence Standard and Basic Competence, three out of five reading versions lacked content validity when judged from the Graduate Competence Standard. The main problem with this research was the absence of a detailed description of the content validation procedure. Rather than presenting details of how it was done through expert judgment, it suddenly showed a table with a percentage of the content validity of the test.

Another study in the same area was conducted by Geribka (2016) and by Jayanti, Husna, & Hidayat (2019). By matching the test specification and the items of the reading test, a crucial step that was done solely by the researcher and by the English teachers themselves, the researchers concluded that the reading test had a high degree of content validity. In fact, a simple eyeballing of the test items and the syllabus could have been made more convincing by the involvement of experts' judgment and computation of Aiken's formula. In other words, the conclusion could have been made more sound if there had been systematic efforts to request the judgment from experts in reading and testing about the match between the test specifications and the items. Furthermore, in the area of construct validation, the crucial steps such as measuring the correlation between the national examination items and a currently established measure of language proficiency or submitting the scores to a factor analysis were not performed.

Another study on content validity of the national examination was carried out by Rukmini and Jumaroh (2015). They examined the content validity of the English national examination which supposedly measured the mastery of the four language skills as required in the so-called *Kurikulum Tingkat Satuan Pendidikan* (KTSP) Curriculum (Curriculum for Educational Unit Level). They found that out of 50 items, 15 were for testing listening, 30 for testing reading, and 4 for testing writing. No item was made for testing speaking. On the grounds of these findings, they concluded that even a high-stake National Examination failed to have sufficient content validity. This conclusion was made after a mere calculation of the proportion of the items compared to the proportion of skills in the standard competences in the curriculum. There was no further analysis of the content of the items to ascertain if they indeed measure the skills stated in the standard competences.

Sabrina's study (2016) focused on the quality of reading texts and comprehension questions in the national examination for senior high school students in the academic year 2015. The researcher addressed the quality of the passages, the match between the test questions and the instructional objectives, the difficulty level, and the distractors of the multiple-choice items that were used for measuring reading comprehension. The researcher did not use a double coder to cross validate the difficulty level and discriminatory power of the items, and then vaguely mentioned the concept of validity in passing. In this way, the study missed an essential procedure for determining those features of the test and therefore can only serve as a preliminary conjecture rather than a valid study.

In brief, it can be concluded from the studies discussed above that content validation studies in Indonesia still have a long way to go before they can turn out high-quality research with sound findings. The fundamental principles of establishing content validation, which include a thorough review of the skills being measured, and expert judgment combined with quantitative analysis (Yusoff, 2019) should be mastered and applied to any endeavor to determine the content validity of a high-stake test such as the national examination. It was revealed that the teachers, too, still need a lot of improvement in their practice of testing their students' English language skills. Part of their problem seems to stem from the entire educational policies. In a study about test construction skills, Rasyid et al. (2024) noted that one type of constraints faced by high school teachers are school-level educational policies which hardly support the development of their skills in testing. The crux of the problem has also been voiced by Lie (2007, p. 1) who stated that 'the underlying policies and practices in the EFL classroom do not always match consistently'.

Quality of Writing Tests

Our search generated only two eligible reports on the quality of tests of writing skills in Indonesian schools. They are summarized in Table 5 below and then elaborated in the section after the table.

Table 5 *Studies on the quality of writing tests*

Writers	Objective	Object of analysis	Comments
Ginting & Saukah (2016)	To describe how 22 high schools in Malang, Indonesia prepared, administered, and scored writing tests.	Teachers of writing from 22 schools.	A proper method for achieving the research objective was used.
Imamyartha & Sulistyono (2017)	To determine whether writing tests in 3 schools meet the criteria for theory-based validity, context-based validity, scoring validity and consequential validity.	The writing tests.	Insufficient methodology for such complex research. objectives.

The foregoing review has revolved around tests of discrete elements and receptive skills like grammar, vocabulary, and reading. Obviously, studies on productive language skills also merit attention. Before the review continues, Table 6 below is presented to summarize the language elements and language skills that were the focus of the studies reviewed so far.

With regard to writing tests, Ginting and Saukah (2016) conducted a study on how 22 high schools in Malang, Indonesia prepared, administered, and scored writing tests. They found that most of the schools (around 76%) did not consult experts for construct validation. The teachers said that such procedure was complicated; consequently, they designed the test themselves. The respondents also did not follow the strict regulation concerning seating arrangement in the test rooms so as to minimize

Table 6 *Language elements and language skills that were measured*

Writers	Language elements and skills measured
Semiun & Luruk (2020)	Vocabulary, grammar, reading comprehension
Hilaldy (2021)	Vocabulary, grammar, reading comprehension
Andriyani & Gustine (2022)	Vocabulary, grammar, reading comprehension
Rohmah (2019)	Vocabulary, grammar, reading comprehension
Sugianto (2020)	Vocabulary, grammar, reading comprehension
Maharany & Putro (2020)	Vocabulary, grammar, reading comprehension
Karim (2021)	Vocabulary, grammar, reading comprehension
Widyaningrum & Prabandari (2013)	Vocabulary, grammar, reading comprehension
Rukmini & Jumaroh (2015)	Vocabulary, grammar, reading comprehension
Geribka (2016)	Reading comprehension
Sabrina (2016)	Reading comprehension
Jayanti, Husna & Hidayat (2019)	Reading comprehension
Ginting & Saukah (2016)	Writing
Imamyartha & Sulisty (2017)	Writing

students' copying each other's work, and did not do cross-scoring, and even did not try out the tests. They reasoned that this procedure was unwieldy and that scoring the essays individually would be more efficient. A few schools, due to the shortage of teachers, simply let every teacher decide the test quality themselves.

The finding above is similar to the findings by Imamyartha and Sulisty (2017), who studied the writing tests of three different schools from the perspective of multifaceted validity. According to the study, the writing tests in the schools failed to meet the criteria for theory-based validity, context-based validity, scoring validity and consequential validity. The researchers also found out that all schools never bothered to establish inter-rater reliability when scoring students' written works, let alone standardizing the raters in the scoring procedure, or consulting experts for the construct validity of their writing tests.

What was largely missing from the study above was a very rigid and complex procedure necessary for determining validity. The researchers based their findings or claims on some simple questions whether the writing tests met all of the requirements to have those types of validity. At the most, this study could serve only as a small-scale preliminary study that could lead to a more complex research investigating the validity of the writing test.

Aside from the weakness of the second study above, it is clear from the situation depicted by the research above that the testing of writing in high schools is riddled with fundamental problems. With almost all of the vital elements in testing writing being ignored by the practitioners in schools, one can only imagine how the validity of such tests has been greatly compromised. Of course, it goes without saying that tests with shaky validity and reliability just cannot be trusted. Teachers and researchers could actually make efforts to understand the fundamental principles of testing writing, which according to (Hughes, 2010) and Lynne (2004) involves careful selection of tasks, consideration of direct and

indirect testing, and the application of steps to ensure validity and reliability which is enhanced by the use of clear rating scales.

The current trends in language testing have been encapsulated in several studies. (Sadeghi & Douglas, 2023) addressed the importance of considerations in a technology-mediated language assessment where language tests are increasingly personalized and real-time feedback can be given. Similarly, (Stošić & Malyuga, n.d.2024) also discussed the increasing use of AI-driven testing instruments which guarantees higher precision and effectiveness. In such context, the importance of traditional testing paradigms like validity and reliability should still be prioritized. A bibliometric analysis by (Yang & Wang, 2025) revealed a trend of writing tests and a concern about the washback effect of language tests. Our review contributes information on how writing tests are conducted in Indonesia and also addresses a similar concern about the washback effect. Thus, when situated in the current trends in language testing, our review are still relevant and even offers a deeper look into the practice of language testing in one developing Southeast Asian country. If testing in schools is being geared up to AI and other forms of technology-assisted testing, care should be taken to ensure that the educators and policy makers have a good grasp of the fundamental principles in language testing.

Conclusion

This paper set out to review previous research on language assessment practices in Indonesian high schools, particularly evaluating the quality of language tests used at both school and national levels. Through the PRISMA framework, 14 research-based journal articles were selected from seven reputable databases, with a focus on assessing their methodological soundness and the implications of their findings. The review indicates that many of the selected studies lack sufficient methodological rigor, which weakens the credibility of their results. Several limitations were consistently observed. First, procedures for establishing validity and reliability were often incomplete or inadequately reported. For instance, while some studies claimed content validity through simple alignment with syllabi, they neglected to involve expert judgment or apply statistical techniques such as Aiken's V coefficient. This omission compromises the validity evidence, as expert evaluations and quantitative agreement indices are essential to confirm whether test items accurately represent the intended constructs. Second, construct validation procedures were frequently absent. Without correlating test scores with established language proficiency measures or conducting factor analyses, these studies provide little assurance that their tests accurately measure the intended skills. Future research should adopt these procedures to enhance the interpretability and trustworthiness of test results. Although a few studies adequately addressed item difficulty, discrimination, and distractor effectiveness, these analyses remain sample-dependent. As such, item characteristics may vary across different groups of test-takers. A more advanced solution, such as employing Item Response Theory (IRT), which accounts for test-taker ability levels independently of the sample, would yield more stable and generalizable findings—an approach yet to be explored in the reviewed studies.

The implications of the findings above concern the number of evaluation studies and the credibility of their results in the eyes of the decision-making parties. It is clear from the small number of the studies that were eventually chosen for analysis that there is a dearth of studies evaluating the quality of language tests in Indonesian schools. Second, the existing studies turned out results which are largely not convincing due to their poor methodological procedures. Clearly, more emphasis and improvement should be made in the area of teaching language testing principles and research methodology. Researchers interested in conducting language testing research should have a more solid and proper knowledge about the necessary procedures for establishing validity, reliability, and authenticity of language tests.

Another equally important implication is that there is a more pressing need to evaluate test quality based on empirical findings. Efforts should be made to increase the teachers' abilities in constructing good language tests, and then to subject their tests to stringent evaluation so improvement can always be made on the basis of the empirical results of the evaluation.

This review has limitations, too. First, the databases from which the articles were obtained were only seven. The articles that were eventually viewed as eligible came mostly from domestic journals. On one hand, this reflects a limitation but on the other hand, it also indicates that no research article on language testing written by Indonesian scholars managed to make their way into international publication. In the future, other researchers could expand the scope by involving a larger number of databases. Another limitation is the fact that only articles written in English were reviewed. The conclusion drawn in this paper, therefore, may have partially missed the profile of reports written in Indonesian language. In the future, another systematic review on language testing could scan larger databases to detect more studies in language testing in Indonesia.

The foregoing discussion points to the existing condition of school English teachers and their know-how of language testing. The overall impression is that the qualities of language tests in schools and the qualities of the research on those testing practices need substantial improvement. Thus, it is the obligation of the ministry of education to ensure that it produces pedagogically sound policies that promote a solid teaching of language testing skills for would-be teachers and encourage good practices in the EFL curriculum, especially in assessment practices.

Abbreviations

PRISMA = Preferred Reporting Items for Systematic reviews and Meta-Analyses

SAGE = Sarah and George (founders of the journal)

ERIC = Education Resources Information Center

JSTOR = Journal Storage

References

- Allanson, P. E., & Notar, C. E. (2019). Writing multiple choice items that are reliable and valid. *American International Journal of Humanities and Social Science*, 5(3), 1–9.
- Andriyani, D., & Gustine, G. G. (2022). Analysis of English formative tests for Junior High School students. *Jurnal Pendidikan Tambusai*, 6(1), 304–311.
- Chapelle, C. (2021). Validity in language assessment. In P. Winke, & T. Brunfaut, *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 11–20). Routledge.
- Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. SAGE.
- Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. SAGE.
- Chapelle, C. A., & Lee, H. (2021). Conceptions of validity. In G. Fulcher, & L. Harding, *The Routledge handbook of language teaching* (pp. 1–25). Routledge.
- Effendi, T., & Suyudi, I. (2016). The impacts of English national examination in Indonesia. *Proceedings of the Ninth International Conference on Applied Linguistics (CONAPLIN)9* (pp. 236–239). Atlantis Press. <https://10.2991/conaplin-16.2017.52>
- Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge University Press.
- Geribka, M. (2016). *A study on the content validity and authenticity of English national examination 2015 for Junior High School*. Universitas Sanata Dharma.
- Ginting, D., & Saukah, A. (2016). Tests of writing in the school examination in upper secondary schools. *SAGE Open*, 6(4). <https://doi.org/10.1177/2158244016673130>

- Haladyna, T. (2022). Creating multiple-choice items for testing student learning. *International Journal of Assessment Tools in Education*, 9(Special Issue), 6–18. <https://doi.org/10.21449/ijate.1196701>
- Hamied, A. F., & Musthafa, B. (2019). Policies in language education in Indonesia. *Indonesian Journal of Applied Linguistics*, 9(2), 308–315. <https://10.17509/ijal.v9i2.20279>
- Hilaldy, R. (2021). Content validity of English summative test. *Journal of Language Learning and Research*, 4(1), 37–41. <https://doi.org/10.22236/jollar.v4i1.7708>
- Hollingworth, L., & Drake, H. M. (2012). *Teach reading, not testing: best practice in an age of accountability*. Corwin Press.
- Hughes, A. (2010). *Testing for language teachers*. Cambridge University Press.
- Imamyartha, D., & Sulisty, G. H. (2017). An evaluation paradox: the issues of test validity in the realm of writing test as the final school examination in the Indonesian senior high school milieu. *Dinamika Ilmu*, 17(1), 1–21. <http://dx.doi.org/10.21093/di.v17i1.585>
- Jayanti, D., Husna, N., & Hidayat, D. N. (2019). The validity and reliability analysis of English National Final Examination for Junior High School. *Voices of English Language Education Society*, 3(2), 127–135.
- Karim, S., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature and Culture*, 6(2), 256–269. <https://10.30659/e.6.2.256-269>
- Komaria, O. (1998). *The history of English teaching in Indonesia*. Applied English Linguistics Program. Jakarta: Graduate School Universitas Katolik Atma Jaya.
- Lie, A. (2007). Education policy and EFL curriculum in Indonesia: between the commitment to competence and the quest for higher test scores. *TEFLIN*, 18(1), 1–14.
- Lynne, P. (n.d.). *Coming to terms: A theory of writing assessment*. https://digitalcommons.usu.edu/usupress_pubs
- Mackey, A., & Gass, S. M. (2022). *Second language research: Methodology and design*. Routledge.
- Maharani, A. U., & Putro, N. H. (2020). Item analysis of English final semester test. *Indonesian Journal of EFL and Linguistics*, 5(2), 491–502.
- Mariappan, V. A., & Osman, K. (2023). A systematic literature review of authentic assessment in K-12 ESL/EFL education. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, 8(5), e002303. <https://doi.org/10.47405/mjssh.v8i5.2303>
- Nababan, P. W. (1991). Language in education: the case of Indonesia. *International Review of Education*, 37, 115–131. <https://doi.org/10.1007/BF00598171>
- Nurdiana. (2020). Language teacher assessment literacy: A current review. *Journal of English Language and Culture*, 11(1), 66–74. <http://dx.doi.org/10.30813/jelc.v11i1.2291>
- Puspawati, I., & Khansa, M. (2024). Developing EFL teachers' language assessment literacy: A systematic literature review on teacher training programs. *Widiati, Utami*, 28(2), 32–42.
- Rasyid, M. N., Alfian, J. D., & Nurwahida. (2024). English language assessment in Indonesian senior high schools: perceived challenges and prospective solutions. *Journal on English as a Foreign Language*, 14(1), 1–25. doi:: <https://doi.org/10.23971/jefl.v14i1.6963>
- Rohmah, N. (2019). Validity and reliability study on teacher-made assessment for English mid-term examination. *Advances in Social Science, Education and Humanities*, 254, 107–110. <https://10.2991/conaplin-18.2019.236>
- Rukmini, D., & Jumaroh. (2015). Does the National Examination of English subject in Indonesia test what should be tested? *Jurnal Pengembangan Humaniora*, 15(3), 136–144.
- Sabrina, A. (2016). A language assessment analysis of English national examination in Indonesia. *EduLite*, 1(2), 177–195.
- Sadeghi, K., & Douglas, D. (2023). Fundamental considerations in technology-mediated language assessment (K. Sadeghi & D. Douglas, Eds.). Routledge.

- Saukah, A., & Cahyono, A. E. (2015). National exam in Indonesia and Its Implications to the Teaching and Learning of English. *Jurnal Penelitian dan Evaluasi Pendidikan*, 19(2), 243–255.
- Semiun, T. T., & Luruk, F. D. (2020). The quality of an English summative test of a public junior high school, Kupang-NTT. *English Language Teaching Educational Journal*, 3(2), 133–141. <https://doi.org/10.12928/eltej.v3i2.2311>
- Stošić, L., & Malyuga, E. N. (n.d.). on-Don, Russian Federation. *Anglisticum Journal* (IJLLIS), 1. <https://aielasconference.com>
- Sugianto, A. (2020). Item analysis of English summative test: EFL teacher-made test. *Indonesian EFL Research & Practice*, 1(1), 35–54.
- Suherman, A. (2022). Exploring language assessment literacy of EFL teachers in the context of Indonesian higher education. *English Review Journal of English Education*, 10(3), 783–792. <https://doi.org/10.25134/erjee.v10i3.6525>
- Sujati, H., Haryani, B. S., Adi, K., & Aprilia, T. (2021). Validity analysis of the elementary school students' tolerance character measurement. In Y. W. Purnomo, & Herwin, *Educational innovation in Society 5.0 era: challenges and opportunities* (pp. 255–259). Routledge.
- Sulistiyo, U., Juwita, M., Hadiyanto, Agustina, P. S., Astini, S., Anwar, K., & Wiryotinoyo, M. (2020). What benefits that students can take? Portraying the implementation of EFL curriculum in Indonesia through phenomenological lens. *Elementary Education Online*, 19(3), 1487–1497. <https://10.17051/ilkonline.2020.731167>
- Susyla, D., & Jaya, S. (2023). Digital assessment in English language teaching (ELT): A systematic literature review, 7(1). <https://journals.unihaz.ac.id/index.php/edu-ling>
- Widyaningrum, F. A., & Prabandari, C. S. (2013). Content validity and authenticity of the 2012 English test in the Senior High School examination. *LLT Journal: A Journal on Language and Language Teaching*, 23–29.
- Yang, Z., & Wang, P. (2025). Current status and research trend of English language assessment: a bibliometric analysis. In *Language testing in Asia* (Vol. 15, Issue 1). Springer. <https://doi.org/10.1186/s40468-024-00317-w>
- Yusoff, M. S. B. (2019). ABC of content validation and content validity index calculation. *Education in Medicine Journal*, 11(2), 49–54. <https://doi.org/10.21315/eimj2019.11.2.6>