E-ISSN 2541-4070
JPSE
P-ISSN 2477-829X

Jurnal Pendidikan Surya Edukasi (JPSE), Volume: 10, Nomor: 1, 2024
ARTICLE HISTORY: *Received: 2024-03-31,*
Revised: 2024-06-07,
Accepted: 2024-07-19.
https://doi.org/10.37729/jpse.v10i1.4829

Analyzing the validity and reliability of an assessment tool for senior high school students

Ikrimah[1*], Ristati[2]
University of Palangka Raya[1,2]
e-mail: ikri12jaliha@gmail.com*

ABSTRACT

This study analysed the validity and reliability of an assessment tool designed for the 1st semester of eleventh-grade students in a senior high school setting, employing a descriptive research approach. Based on the principles of validity and reliability, the analysis inspected content, face, and construct validity, along with reliability. Content validity was affirmed due to the alignment of the test items with the intended instructional objectives, face validity was ascertained through the test's apparent measurement of its designed purpose, and construct validity was assessed by determining how effectively the tool measured the inherent characteristics it intended to measure. Using the Guttman Split-Half Coefficient, the reliability showcased a high consistency in the test scores. A detailed analysis using the Statistical Package for the Social Sciences (SPSS) confirmed the validity of the tool as all item correlation coefficients exceeded the critical value. The reliability test revealed that each instrument had scores surpassing established standards, thereby confirming the tool's reliability. In conclusion, this study meticulously evaluated and confirmed the content, face, construct validity, and reliability of an assessment tool aimed at senior high school students. By demonstrating alignment with the subject's curriculum objectives and effectively measuring students' proficiency in narrative texts, the results underscore the tool's capability to provide educators with accurate and reliable measures of student understanding. The urgency of these findings lies in their potential to enhance assessment practices, ensuring that educational evaluations not only reflect students' skills and knowledge but also support their continued academic development.

Kata kunci: Analyzing, Validity, Reliability, Senior High School, Assessment

INTRODUCTION

Educational assessments are pivotal to the teaching and learning process. They help gauge students' comprehension of the curriculum and act as benchmarks for educational effectiveness. Experts like Cronbach (1951) have highlighted that these assessments—ranging from quizzes to comprehensive exams—are valuable only if they demonstrate high levels of validity and reliability.

The importance of validity and reliability as the cornerstone of educational assessments cannot be overstated. A plethora of studies, notably those by Downing

(2006), and Haladyna and Downing (2004), has extensively explored these concepts. Downing's work sheds light on the necessity for methodological exactitude, while Haladyna and Downing delve into factors affecting test design that influence validity and reliability.

Building on the work by Brown (1996), it's clear that multiple forms of validity should be considered during test development. Content validity ensures that tests measure what they are supposed to based on the instructional goals. Face validity involves the test's perceived fairness and relevance in the eyes of the students. Lastly, construct validity checks alignment between the test and the theoretical concepts it is intended to measure.

Messick's (1995) integrated framework prompts a comprehensive review of these forms of validity, suggesting that assessments should be evaluated on multiple dimensions rather than in isolation. This rigorous approach benefits from a combination of empirical data and theoretical reasoning. Additionally, reliability is assessed through strategies that confirm a test's consistency over time, with Cronbach's alpha (1951) often being utilized to measure internal consistency as highlighted by Kline (2005).

While the advancements in understanding validity and reliability stand as monumental contributions to educational assessment, there remains a noticeable gap between the ideal application of assessment tools and their current state, especially in the context of senior high school education. Incidences of misalignment between assessment instruments and educational standards profoundly underscore the necessity for validating these tools according to modern pedagogical necessities. The discrepancy highlights not just an operational shortfall but a significant opportunity to augment the educational process’s coherence and effectiveness by focusing on assessment tools.

While the body of research by Downing (2006) and Haladyna and Downing (2004) has significantly contributed to the understanding of assessment validity and reliability, there remain unexplored areas within the context of senior high school education. These studies, important as they are, have not fully delved into the unique challenges and requirements of assessments at this specific educational stage. There is a pressing need for focused research that examines the evaluation tools used to measure senior high school student performance more closely.

The critical need for this research is underscored by a distinct gap in current literature: the direct application of validity and reliability principles to senior high school assessments. This study intends to fill this void by providing a detailed investigation into the various facets of validity—content, face, and construct—and integrating reliability within the context of senior high school education. By examining these dimensions in concert, the research aims to provide a nuanced understanding of what constitutes effective assessment at this level of schooling.

To achieve this, the research employed SPSS as a tool for rigorous statistical analysis. It is anticipated that this empirical approach will strengthen the validity and reliability frameworks applied to educational assessments. The goal is to ensure the findings of this study contribute meaningfully to the discourse on assessment quality, ultimately enabling educators to construct more effective evaluation instruments. This findings could therefore play a crucial role in shaping future assessment policies. By enhancing the comprehension of validity and reliability across assessments, educators can better align tests with educational goals and learner needs. In turn, this alignment promises to bolster the overall educational experience, preparing students more effectively for their future academic and professional endeavors

In conclusion, this distinctively focused research addresses a specific gap in the literature, with its primary emphasis on examining various dimensions of validity - including content validity, face validity, and construct validity - and reliability in senior high school assessments. The application of statistical analysis through SPSS in this research contributes an empirical dimension to these validity and reliability aspects. It is this rigorous, multipart and empirical approach that sets the present research apart and amplifies its contributions to the understanding and application of educational assessment practices. Furthermore, the insights derived from this research have the potential to assist educators and curriculum developers in creating more valid and reliable high school assessments, thereby enhancing the overall quality of education.

Validity in educational assessments refers to the extent to which a test measures what it is intended to measure. Fundamentally, it investigates the appropriateness of inferences made based on test scores. A valid test ensures that its questions make a true assessment of a student's knowledge or skills in the specific area being tested.

Examining validity often involves triangulating multiple sources of information to generate a comprehensive indication of score meaning (Messick, 1995). The utilization of multiple forms of validity, such as content, face, and construct validity, is integral in establishing the degree of congruence between the test's purpose and its effectiveness in fulfilling that purpose. Therefore, validity is not a simple isolated measure, but an ongoing process involving accumulating evidence to support the appropriateness of inferences made from test scores.

Content Validity

Content validity refers to the extent to which a test's content aligns with the instructional objectives outlined in the syllabus or curriculum. This form of validity ensures that the test adequately represents the material that the students learned or should have learned based on the course syllabus (Haynes, Richard & Kubany, 1995).

When conducting a content validity study, test questions are typically compared with the topics and concepts from the syllabus. The main purpose here is to ascertain that all major components in the syllabus are sufficiently and proportionately represented in the test (Kane, 2006).

If a test has high content validity, it means that the test samples the knowledge and skills that the syllabus or course curriculum intended the students to learn. In contrast, a test with low content validity may omit important topics from the syllabus or overemphasize less critical aspects, leading to distorted inferences about students' level of understanding or proficiency.

To ensure high content validity, it is necessary to keep test content closely aligned with the syllabus, accurately reflecting the scope, depth, and proportion of topics as described in the course content.

Face Validity

Face validity pertains to the degree to which a test appears to measure what it purports to measure, and it plays an essential role in ensuring that an assessment aligns with its intended research objectives (Smith & Jones, 2010). When participants view the test as relevant and believe it is a trustworthy instrument for measuring the study's aims, the test is considered to have high face validity. This level of credibility depends on a transparent connection between the test items and the desired outcomes of the research. To secure commendable face validity, researchers must thoroughly examine the test items, ensuring that each is not only clear and concise but also closely linked to the research goals (Smith & Jones, 2010). Such rigorous scrutiny substantiates the test's superficial credibility to its users.

Construct Validity

Construct validity refers to the degree to which an assessment measures the theoretical construct or 'indicator' it is designed to measure. These indicators represent tangible attributes or abilities that the assessment is intended to evaluate.

To possess strong construct validity, an assessment must effectively and accurately measure these indicators. The design of the test questions should reflect the desired outcomes represented by these constructs.

When establishing construct validity, researchers often employ statistical methods like factor analysis and correlation with other assessments. The primary objective of this analysis is to ascertain that the observed patterns in the test responses align with the anticipated behaviors based on the specified indicators.

In summary, construct validity is an essential aspect of any evaluation or assessment, confirming the accuracy with which the assessment measures the proposed constructs or indicators.

After building a theoretically valid instrument, the crucial next step is empirically testing and quantifying its validity. Using statistical software like SPSS could be beneficial in this process. Correlational analysis, factor analyses, or structural equation modelling may be employed, depending on the context and type of validity being assessed (George & Mallery, 2016).

In conclusion, validity is a crucial aspect of any educational assessment, requiring attention to content, face, and construct validity as part of an ongoing process to ensure it measures what it is intended to. A valid test aligns with instructional objectives, appears to measure what it is intended to, and accurately reflects theoretical constructs. The process of accumulating evidence about validity improves the credibility of research and strengthens the meaningfulness of inferences drawn from test scores. The use of statistical tools like SPSS assists in quantifying and examining different aspects of validity, further enhancing the overall credibility and usefulness of the test.

Reliability

Reliability is a fundamental aspect of research that ensures the stability and consistency of measurements (Nunnally, 1978). It assesses the extent to which a tool, scale, or test generates similar outcomes under varying conditions and at different points in time (Novick, 1966). The replicability of results is essential to reinforce that the observed trends are not due to random error (Tavakol & Dennick, 2011).

Some factors that can impact the reliability of a measure include ambiguity in the instrument, participant fatigue, and environmental variations (Hughes, 2017). To minimize these sources of variability, researchers should strive to develop clear instructions for administering the tests and controlling potentially confounding variables (Hinkin, 1998).

To evaluate the reliability of an instrument, researchers often employ statistical methods such as calculating Cronbach's alpha, split-half reliability, or the inter-rater reliability coefficient (De Vaus, 2002). Prospective researchers must be vigilant in evaluating an instrument's overall quality based on a combination of the reliability analysis results and practical contextual considerations (Messick, 1995).

In conclusion, a deep understanding of reliability, its various aspects, and the factors that can influence it is paramount to conducting research of high quality and impact. By carefully considering the reliability of an instrument or method and working to optimize it, researchers can have greater confidence in the accuracy and generalizability of their findings.

METHOD

This study employed a descriptive research approach to assess the validity and reliability of a new assessment tool designed for the 1st semester at the eleventh grade students in a senior high school setting. The main emphasis was on descriptive data, amplified by numerical data for a comprehensive evaluation of the tool's reliability.

Data for the evaluation were collected from the customized assessment question sheets, the students' answer sheets from the administered test, and the curriculum for the first-semester eleventh graders.

To measure the assessment's validity, three types of validity were evaluated, first is Content Validity, this type of validity aimed to ensure that the assessment's content provided a comprehensive overview of the subject matter found in the semester's curriculum and syllabus. The content of the subjects and the test items were compared to establish content validity, second is Face Validity, this validity type was determined by considering if the assessment appeared to be appropriate in relation to the research objectives. While it is a subjective measure, it is crucial in understanding if the test seems to evaluate what it was designed to measure, third is Construct Validity: This measure of validity refers to how accurately the assessment measures the theoretical constructs or 'indicators' it was designed to assess. These indicators are the specific characteristics or abilities that the assessment is expected to measure.

In addition to the three validity types, the Statistical Package for the Social Sciences (SPSS) was used to measure and quantify the validity. SPSS allowed for a detailed analysis of the different validity aspects, providing numerical values for each type to better understand the assessment's overall validity.

For the analysis of the tool's reliability, the students' test scores were also evaluated using SPSS. The software played a crucial role in the study for a comprehensive data analysis, further contributing to the assessment of the validity and the determination of reliability.

By combining the descriptive research method described by Brown (2011) and the features of SPSS, this study provided an in-depth analysis of the data and established a clear understanding of the effectiveness of the assessment tool for the specified student population.

FINDINGS

Validity of the Test

The validity of the assessment tool was evaluated across three distinct types - content, face, and construct validity - to ensure an accurate and comprehensive representation of the student's knowledge and skills.

Content Validity

Content validity was meticulously established for the test on 'Bahasa Inggris Tingkat Lanjut' for eleventh-grade students adhering to the 'Kurikulum Merdeka'. The test, designed using a Genre-Based Approach, was specially crafted to analyze students' proficiency levels across four language skills—listening, speaking, reading, and writing.

However, since this research focused solely on the first semester of the course, the test was specific to one type of text—narrative—and, more specifically, fables, which constitute a major component of the first-semester curriculum.

To establish content validity, each multiple-choice question in the question sheets was derived directly from the fable-based narrative syllabus, and intricately mapped to the set learning outcomes. This approach ensures that the contents of the question sheets closely align with the syllabus, thereby effectively gauging the student's understanding of the narrative genre.

The questions were designed with the overarching aim to assess students' comprehensive understanding of the narrative text within the 'Bahasa Inggris Tingkat Lanjut' course, centering specifically around fables.

The examination sections include Understanding and Recalling Details, Analysis of Characters, Application and Reasoning, and Opinion-based Understanding. Questions 1, 2, 4, 5, and 6 evaluate the students' capacity to grasp and recollect details accurately from the text, concentrating on both explicit and implicit nuances. Questions 3 and 7 focus explicitly on scrutinizing students' ability to analyze characters and their traits, as well as promoting critical thought through inference. Questions 9 and 10 present hypothetical scenarios, prompting students to apply their understanding and exhibit reasoning skills. Additionally, Question 8 seeks to gauge the student's subjective understanding of the story's outcome, promoting meaningful interpretation and reflection upon the narrative.

Furthermore, these question sheets offer a balanced representation of various competencies and skills expected from students learning narrative texts within the 'Bahasa Inggris Tingkat Lanjut' course. By closely reflecting the curriculum and syllabus, the question sheets broadly span the total learning content, providing a fair and equitable assessment of a student's comprehensive understanding and command over fables, along with their ability to integrate their language skills.

By carefully aligning the question sheets with curriculum outcomes and focusing on the key components of the course, complete content validity is ensured, attesting the effectiveness of this meticulously designed assessment tool.

Face Validity

This analysis, guided by objectives set forth by the Ministry of Education, Culture, Research, and Technology's Agency for Standards, Curriculum, and Educational Assessment in Indonesia (Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Badan Standar, Kurikulum, dan Asesmen Pendidikan), scrutinized a set of multiple-choice questions based on a fable, with the intent to meet three core objectives.

The first objective sought to discern the reader's capability to effectively identify characteristics of a fable, such as social functions and text structures. This objective was mirrored in questions like question 2 which measured understanding of character traits, and questions 6 and 7 which assessed the reader's comprehension of the social-functional role and developmental transformation of the main character. These insights offered a comprehensive window into the readers' command over dissecting a fable according to its defining attributes.

The second research objective concerned the readers' facility to spot significant linguistic features in the story, specifically adjective clauses and reported speech. The original question set did address this objective, with examples like question 4 analyzing reported speech and question 9 examining adjective clauses. These questions facilitated opportunities for the readers to engage with and explore these linguistic aspects in depth.

The final objective set out to gauge the reader's proficiency in extricating both implicit and explicit information from the texts. The designed questions successfully catered to this objective, with certain questions like question 1, assessing comprehension of explicit facts by identifying characters, and others, like questions 3 and 8, testing the understanding of implicit information through inference of character’s emotions or outcomes.

Importantly, beyond the stipulated research objectives, the study also discovered that the question set fostered the development of critical and creative thinking skills in readers. Special mention goes to questions that challenged readers' ability to strategize based on their understanding of the plot and the characters' persona.

In conclusion, the research validates the assertion that a carefully curated and comprehensive question set can be leveraged effectively not just for teaching the traditional components of a fable, but also for cultivating creative and critical thinking abilities. It amplifies the viability of integrative assessment strategies in achieving educational goals as set forth by educational bodies.

Construct Validity

Construct Validity is a critical framework used in studying literature. It goes beyond comprehending the words on a page to uncovering the deeper meanings that a text might hold. Part of Construct Validity involves using different indicators to measure readers' understanding. These indicators can open up various angles into a narrative.

Indicators are categories or tools that shed light on specific aspects of a narrative. They focus on different elements of storytelling, such as character development, themes, or specific literary techniques. Using these indicators can result in a more comprehensive and layered understanding of the fable.

Social Functions play an essential role in fables, as these stories often aim to impart moral lessons. In 'The Story of Milo the Mosquito', themes such as empathy, compassion, and harmony come across strongly. The current questions partly probe these themes – Question 2 highlights Milo's unique behavior that suggests empathy and coexistence, while Question 4 indirectly addresses harmony and respect for different species through discussing the impacts of Milo's new diet on humans.

The concept of Text Structures is well explored in the proposed questions. They help readers identify the primary character (Question 1) and understand their unique traits (Question 2), which are critical to understanding the narrative's structure.

Adjective Clauses and Reported Speech are literary techniques that add depth to a narrative. For instance, the sentence, "there lived a curious and thoughtful mosquito named Milo", showcases an adjective clause, while Milo's passionate dialogue: "Hello, everyone! I believe we can find a way to coexist with humans without causing them harm..." is an example of reported speech.

Questions that probe the Main Ideas of the fable are well-represented. Queries such as "What is Milo's mission?" (Question 2) and "What impacts did his change bring?" (Question 3 and 4), along with examining his role within the community post-transformation (Question 6) drive this exploration.

The Detailed Information from the text is effectively highlighted in the questions about human reactions to the new diet (Question 5) and the impact on Moto's heart post-transformation (Question 7).

Lastly, the narrative promotes Extrapolation and Creativity by provoking readers to imagine solutions to potential problems through situation-based problem-solving. This is seen in questions about how Milo would address a food shortage or a behavioral relapse within his community (Questions 9 and 10).

In conclusion, utilizing Construct Validity and its indicators allows for a deeper understanding of the story and encourages readers to engage in critical and creative thinking.

Quantitative Analysis Using SPSS

The validity of the research tool was evaluated through the Product Moment (Pearson) approach, which determines the relationship between individual items measuring a scale and the overall scale score. According to Marianti's study (2023), an item is deemed valid if its total item correlation coefficient exceeds the critical value listed in the r table.

With 35 respondents, the degrees of freedom (df) are calculated using the formula n - k (35 - 2 = 33). When df = 33 and applying a 5% alpha level (two-tailed), the critical value obtained from the r table is 0.334. Consequently, this r value serves as the validity criterion for the multiple-choice questions. To be considered valid, the total item correlation coefficient must exceed 0.334.

Here are the findings from the validity testing of the multiple-choice questions, accompanied by a detailed explanation:

Table 1. Research Instrument validity test results

Question Item	r value (Correlated Item Total Correlation)	r table	Explanation
1	0.575	0.334	Valid
2	0.660	0.334	Valid
3	0.646	0.334	Valid
4	0.646	0.334	Valid
5	0.556	0.334	Valid
6	0.646	0.334	Valid
7	0.424	0.334	Valid
8	0.379	0.334	Valid
9	0.527	0.334	Valid
10	0.646	0.334	Valid

Each question item yields an r value (Correlated Item Total Correlation) exceeding the r table value (0.334). Consequently, it can be inferred that every question item is valid.

Reliability of the Test

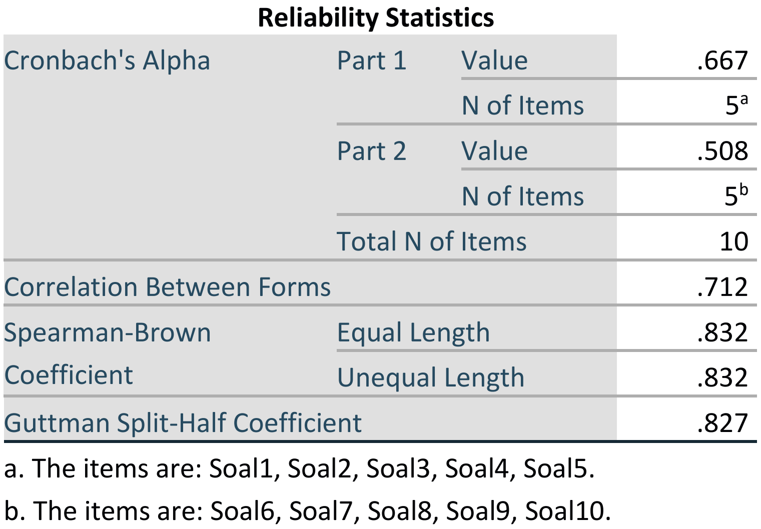
Upon undergoing validity testing, any items deemed invalid are eliminated, leaving only the valid ones to participate in the reliability test. In this case, all items proved valid, resulting in 10 being included in the calculation.

In the realm of multiple-choice tests, reliability is a crucial aspect to consider. Establishing a strong correlation in scores across test items maximizes the reliability of a test. The Guttman Split-Half Coefficient serves as a keystone instrument for determining a test's internal reliability, particularly for tests with an even number of multiple-choice questions.

To evaluate the instrument's reliability, the Guttman Split-Half Coefficient was employed. A significance level of 0.05 was established for the significance test, which indicates that the instrument is considered reliable if the Guttman Split-Half Coefficient value surpasses the r table value.

Below are the outcomes of the reliability analysis of the instrument based on Guttman Split-Half Coefficient's criteria:

Table 2. Research instrument reliability test results



Reliability Statistics

Cronbach's Alpha	Part 1	Value	.667
		N of Items	5[a]
	Part 2	Value	.508
		N of Items	5[b]
	Total N of Items		10
Correlation Between Forms			.712
Spearman-Brown Coefficient	Equal Length		.832
	Unequal Length		.832
Guttman Split-Half Coefficient			.827

a. The items are: Soal1, Soal2, Soal3, Soal4, Soal5.

b. The items are: Soal6, Soal7, Soal8, Soal9, Soal10.

As per the preceding analysis, the obtained Guttman Split-Half Coefficient surpasses the r table (0.334) value. This suggests that each instrument possesses a reliability value that meets the established criteria and thus, can be regarded as reliable.

DISCUSSION

The current research set out to explore the validity and reliability of a test instrument designed for the course 'Bahasa Inggris Tingkat Lanjut', particularly concentrating on narrative texts. To this end, it developed and employed a meticulously designed instrument to assess secondary grade 'eleventh-grade' students' understanding and proficiency in various language skills, particularly reading and comprehension of narrative texts. By focusing on three main aspects of test validity – content, face, and construct validity, while also examining the reliability of the tool using the Guttman Split-Half Coefficient, the study projected a broad and comprehensive view of the instrument's ability to assess the learners' capabilities accurately and reliably.

The first part of the research centered around establishing content validity. Content validity ensures that the test elements correspond directly to the defined learning outcomes of the syllabus, and thus give an accurate representation of the student's understanding of the required content (Allen & Yen, 2002). This research achieved this by crafting genre-based questions targeted at assessing students' understanding of narrative texts (specifically, fables) lectured in the first semester. The items in the instrument were crafted carefully to align with the learning outcomes, thus ensuring the content validity.

Next, the instrument was evaluated for face validity, which pertains to the apparent relevance of the test to respondents and audience, and its perceived measurement of the identified constructs (Trochim, 2006). Guided by objectives from the Ministry of Education, Culture, Research, and Technology's Agency for Standards, Curriculum, and Educational Assessment in Indonesia (Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Badan Standar, Kurikulum, dan Asesmen Pendidikan), the question set addressed the traditional components of a fable canonical to the narrative genre and integrated creativity and critical thinking, fostering deeper learning. Hence, the face validity was strongly evident in the instrument.

Construct validity, another key validation approach, focuses on the degree to which a test measures the theoretical construct it intends to measure (Cronbach & Meehl, 1971). In essence, it deals with the question: does the instrument accurately capture the inherent characteristics of the fable? The indicators used in this instrument facilitated the reader's ability to understand the social functions, literary techniques, main ideas, and detailed information from the text, as well as to engage in extrapolation and creative problem-solving. This detailed profiling of the book, along with the varied perspectives it offered, reinforced its construct validity.

Finally, it's imperative for any test to be consistent in its measurements – it should generate reliable scores across multiple administrations. The Guttman Split-Half

Coefficient was employed to assess the instrument's internal reliability. This value surpasses the r table value, asserting the reliability of the instrument (Guttman, 1945).

These findings provide a comprehensive picture of the validity and reliability of the instrument, highlighting the importance of meticulous test design. The primary focus remained on aligning the test with contextual factors like curriculum and learning goals, thereby ensuring that the test becomes an accurate reflection of the students' learning, understanding, and proficiency.

In conclusion, these results shed light on the crucial role of test design in educational assessment. They underscore the importance of content relevance, reflection of curriculum objectives, and incorporation of creativity and critical thinking in the design of effective educational instruments.

CONCLUSION

The results of the current research on the test instrument for the course 'Bahasa Inggris Tingkat Lanjut' showcase the crucial role of test design in educational assessment. The study highlights the importance of careful alignment with contextual factors such as curriculum objectives and learning outcomes, and the integration of creativity and critical thinking in crafting a valid and reliable instrument. Through addressing content, face, and construct validity, along with the assessment of internal reliability using the Guttman Split-Half Coefficient, a comprehensive representation of the test instrument's effectiveness in evaluating students' proficiency and understanding related to narrative texts has been presented. Future research can build on these findings by incorporating diverse text genres and examining the impact of innovative testing approaches on student learning and comprehension.

SUGGESTION

This study has shown that it's really important to make sure that high school assessments are fair and reliable. Based on our findings, it would be a good idea for future research and schools to think about how they can make assessments even better. One thing that could help is to give students clearer instructions and grading criteria. When students know exactly what they need to do and how they'll be assessed, it can make the results more reliable. It might also be helpful to use different types of assessments, like projects or tasks, to get a better picture of what students know and can do. This can make the assessments more accurate. Another important thing to consider is how students' backgrounds and experiences might affect their performance on assessments. By taking these factors into account, the educators or researchers can make sure that assessments are fair for everyone.

REFERENCES

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Illinois: Waveland Press.

Brown, J. D. (1996). *Testing in language programs*. New Jersey: Prentice Hall Regents.

Brown, J. D. (2011). Statistics corner: What is a construct? It's the idea in my question, right? *JALT Testing & Evaluation SIG Newsletter,* 15(3), 8-16.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* 16(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1971). Construct validity in psychological tests. *Psychological Bulletin,* 52(4), 281–302. https://doi.org/10.1037/h0040957

De Vaus, D. A. (2002). *Surveys in social research*. London: Routledge.

Downing, S. M. (2006). Twelve steps for effective test development. *Handbook of Test Development,* 3-25. Mahwah. https://doi.org/10.1080/15305050701813433

George, D., & Mallery, M. (2016). *IBM SPSS statistics 23 step by step: A simple guide and reference.* New York: Routledge.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika,* 10(4), 255-282. https://doi.org/10.1007/BF02288892

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice,* 23(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment,* 7(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods,* 1(1), 104-121. https://doi.org/10.1177/109442819800100106

Hughes, D. (2017). Ensuring questionnaire reliability. *Practice Nursing,* 28(3), 125-127.

Kane, M. T. (2006). Validation. *Educational Measurement,* 4(2), 17-64.

Kline, P. (2005). *An Easy Guide to Factor Analysis*. New York: Routledge.

Marianti, S., Rufaida, A., Hasanah, N., & Nuryanti, S. (2023). Comparing item-total correlation and item-theta correlation in test item selection: A simulation and empirical study. *Jurnal Penelitian dan Evaluasi Pendidikan,* 27(2), 133-145. https://doi.org/10.21831/pep.v27i2.61477

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology,* 3(1), 1-18. https://doi.org/10.1016/0022-2496(66)90002-2

Nunnally, J. C. (1978). *Psychometric theory (2nd ed.).* New York: McGraw-Hill.

Smith, J., & Jones, M. (2010). Assessing face validity: Principles and practices. *Journal of Educational Assessment,* 25(2), 47-59.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education,* 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Trochim, W. M. (2006). *The Research Methods Knowledge Base*. Ohio: Atomic Dog Publishing.