JURNAL KEPEMIMPINAN & PENGURUSAN SEKOLAH
Homepage : https://ejurnal.stkip-pessel.ac.id/index.php/kp
Email : jkps.stkippessel@gmail.com

p-ISSN : 2502-6445 ; e-ISSN : 2502-6437
Vol. 11, No. 3, June 2026
Page 2026-2036

Jurnal Kepemimpinan & Pengurusan Sekolah

VALIDITY AND RELIABILITY ANALYSIS OF MULTIPLE CHOICE QUESTIONS ON THE QURAN AND HADITH IN THE SUBJECTS OF DA'WAH, KHUTBAH, AND TABLIG FOR GRADE XI AT MAN 1 MANDAILING NATAL

Febriansyah Siregar[1], Sri Wahyuni Khairani[2], Syamsiah Depalina Siregar[4]
[1,2,3,4] Sekolah Tinggi Agama Islam Negeri Mandailing Natal, Indonesia

Email: claaymore31@gmail.com

DOI: https://doi.org/10.34125/jkps.v11i3.2626

Sections Info

Article history:
Submitted: 23 March 2026
Final Revised: 11 April 2026
Accepted: 16 May 2026
Published: 21 June 2026

Keywords:
Validity,
Reliability,
Difficulty Level,
Discrimination Index
Learning Evaluation

ABSTRAK

This study aims to analyze the validity, reliability, difficulty level, and discriminatory power of multiple-choice questions on the material of Da'wah, Sermons, and Tablig for class XI at MAN 1 Mandailing Natal in the 2025/2026 academic year. The type of research used is descriptive quantitative, with 20 class XI students as subjects. Data were collected through a learning outcome test instrument, which was then analyzed using Microsoft Excel. Validity analysis was carried out using the Pearson Product Moment correlation formula, while reliability was calculated using the Cronbach Alpha formula. The results showed that out of 10 questions, there were 7 valid questions and 3 invalid questions. The reliability value (R_{11}) showed a very high level of consistency, so the instrument can be categorized as reliable. Analysis of the level of difficulty showed 30% of the questions were easy, 50% were moderate, and 20% were difficult, while the discriminatory power analysis showed 60% of the questions were categorized as good, 10% were sufficient, and 30% were poor. Overall, the test instrument is considered to be of good quality and suitable for use as an evaluation tool for Islamic Religious Education learning, although several items need to be revised to achieve more optimal construct suitability

ABSTRAK

Penelitian ini bertujuan untuk menganalisis validitas, reliabilitas, tingkat kesulitan, dan daya pembeda soal pilihan ganda pada materi Da'wah, Khotbah, dan Tablig untuk kelas XI di MAN 1 Mandailing Natal pada tahun ajaran 2025/2026. Jenis penelitian yang digunakan adalah deskriptif kuantitatif, dengan 20 siswa kelas XI sebagai subjek. Data dikumpulkan melalui instrumen tes hasil belajar, yang kemudian dianalisis menggunakan Microsoft Excel. Analisis validitas dilakukan dengan menggunakan rumus korelasi Pearson Product Moment, sedangkan reliabilitas dihitung menggunakan rumus Cronbach Alpha. Hasil penelitian menunjukkan bahwa dari 10 soal terdapat 7 soal yang valid dan 3 soal yang tidak valid. Nilai reliabilitas (R_{11}) menunjukkan tingkat konsistensi yang sangat tinggi, sehingga instrumen dapat dikategorikan reliabel. Analisis tingkat kesulitan menunjukkan 30% soal tergolong mudah, 50% sedang, dan 20% sulit, sedangkan analisis daya pembeda menunjukkan 60% dari pertanyaan-pertanyaan dikategorikan sebagai baik, 10% cukup, dan 30% kurang. Secara keseluruhan, instrumen tes dianggap memiliki kualitas baik dan layak digunakan sebagai alat evaluasi pembelajaran Pendidikan Agama Islam, meskipun beberapa item perlu direvisi untuk mencapai kesesuaian konstruk yang lebih optimal.

Kata kunci: *Validitas, Keandalan, Tingkat Kesulitan, Indeks Diskriminasi, Evaluasi Pembelajaran*

INTRODUCTION

In educational institutions, especially formal ones, the most important thing is learning. Evaluation is an important part of learning as a final assessment at the end of the semester, which is one of the most common types of evaluation conducted by educational institutions. According to Syahputra et al. (2020), evaluation is a systematic process used to assess and determine the level of student achievement against predetermined learning objectives. Evaluation is also closely related to assessment and measurement. Meanwhile, Nitko & Brookhart (2014) emphasize that good evaluation must be able to provide relevant information for educational decision-making, whether for diagnostic, remedial, or student learning outcome assessment purposes. Therefore, evaluation plays an important role in ensuring the quality of education.

Learning evaluation is closely related to two other main concepts, namely measurement and assessment. These three concepts have a hierarchical relationship. According to Arikunto (2018), measurement produces quantitative data, assessment interprets this data qualitatively, while evaluation is the process of summarizing the results of assessment to make educational decisions. This is in line with Popham's (2017) opinion that evaluation is "the process of making judgments based on measurement results," which is the process of making decisions based on the results of objective measurements. Meanwhile, according to Anderson (2019), measurement serves as the scientific basis for determining the value (judgment) used in educational evaluation. Evaluation is preceded by assessment, while assessment is preceded by measurement. Tests are a way of indirectly measuring a person's abilities through their responses to stimuli or questions they have answered (Mardapi 2008).

In school education, one form of assessment that is often used is summative testing. These tests are conducted at the end of a learning period or semester to measure overall learning outcomes. Alfani (2022) explains that summative testing aims to provide a comprehensive picture of the extent to which students understand, master, and are able to apply the material that has been taught. The most common form of test used in summative tests is multiple-choice questions, as they can cover many ability indicators with efficient completion time and results that are easy to interpret. According to Haladyna & Rodriguez (2013), multiple-choice tests have advantages in terms of objectivity, time efficiency, and the ability to measure various cognitive domains, ranging from basic understanding to higher-order thinking skills (HOTS). However, the quality of multiple-choice tests is highly dependent on the extent to which the items meet quality analysis standards, such as validity, reliability, difficulty level, and discriminating power (Lubis, 2026).

The quality of a test instrument can be determined through item analysis. This analysis aims to assess the extent to which an item can function properly in measuring students' abilities (Lubis, 2024). According to Novia et al. (2020), a quality item is one that has a high level of validity and reliability, as well as a balanced difficulty index and discriminating power. Astuti (2018) adds that item analysis is also used to identify factors that reduce the credibility of measurement results. Meanwhile, Ida & Musyarofah (2021) state that the results of item analysis can be used as a basis for improving and revising poor items so that the test is more representative. This opinion is in line with Crocker & Algina (2015), who emphasize that validity and reliability tests are two main aspects in developing quality measuring instruments. Without these two aspects, measurement results cannot be trusted as a representation of the actual abilities of students.

Furthermore, validity and reliability are two main criteria in assessing the suitability

of an evaluation instrument. Hayati (2016) explains that validity indicates the extent to which a test actually measures what it is supposed to measure, while reliability relates to the level of consistency of measurement results. A valid instrument is not necessarily reliable, but a reliable instrument will support the validity of the measurement. George & Mallery (2018) add that in the context of psychometric testing, good reliability is usually in the range of 0.70–0.90, while values above 0.90 indicate very high consistency. Therefore, before an instrument is used as a learning outcome measurement tool, it is necessary to test its validity, reliability, level of difficulty, and discriminating power. Based on these considerations, this study is important to analyze the quality of multiple-choice questions on the materials of Da'wah, Khutbah, and Tablig for grade XI at MN 1 Mandailing Natal, to ensure that the instrument is suitable and accurate for use in evaluating learning outcomes of the Quran and Hadith.

METHODS

A quantitative approach with a descriptive type was used in this study. The descriptive quantitative approach aims to describe the characteristics of a population or a particular phenomenon without testing the relationship between variables (Creswell & Creswell, 2017). The data obtained was classified as quantitative data, namely the formative test instruments for Da'wah, Khutbah, and Tablig materials. The data collected consisted of test sheets, answer sheets, and answer keys. The participants in this study were 20 grade XI students at MAN 1 Mandailing Natal in the 2025/2026 academic year in Mandailing Natal Regency. The evaluation was carried out on the test instrument items. Data analysis was performed using Ms. Excel. The analysis techniques applied consisted of validity testing, reliability testing, difficulty level testing, item discrimination analysis, and distractor analysis. The analysis results were represented in tables and interpreted descriptively.

RESULT AND DISCUSSION

A. Validity Test Result

Validity testing is a process to determine the extent to which an instrument or item Validity testing is a process to determine the extent to which an instrument or item is capable of measuring what it is supposed to measure. The validity of an item is fulfilled if it has a significant correlation with the total score, meaning that the score on the item is in line with the increase or decrease in the overall score of the students. According to Creswell (2018), validity relates to "the degree to which evidence and theory support the interpretations of test scores." This means that the validity test results not only show correlation figures, but also empirical and theoretical evidence that the instrument is appropriate for the measurement objectives.

$$r_{hitung} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{\{N\Sigma X^2 - (\Sigma X)^2\}\{N\Sigma Y - (\Sigma Y)^2\}}}$$

Keterangan :

r.calculate: correlation coefficient

N : many students

X : score on the test item

Y : total score

In this study, validity testing was conducted on the Al-Quran Hadith test instrument on the materials of Da'wah, Khutbah, and Tablig that had been given to 11th grade students at MAN 1 Mandailing Natal in the 2025/2026 academic year. This analysis was carried out using Microsoft Excel, using the Pearson Product Moment correlation formula shown in Table 1.

Table 1. Validity Test Results for the Test Instrument

NO	Student Name	Question Item (X)										Total Score (Y)
		1	2	3	4	5	6	7	8	9	10	
1	ADE	1	1	1	1	1	1	0	1	0	1	8
2	ADIZ	1	1	1	0	0	1	0	1	0	1	6
3	FAJRI	1	0	1	0	0	0	0	1	1	0	4
4	AJIRNI	1	1	1	1	1	1	0	1	0	1	8
5	ALIFAH	1	1	1	0	0	0	0	1	0	0	4
6	ALYA	1	1	1	1	1	1	0	1	0	1	8
7	ANNISA	1	0	1	0	0	0	0	1	0	1	4
8	DIAH	1	0	1	0	1	0	0	1	0	0	4
9	EKA RAMDANI	1	0	1	0	0	1	1	1	0	0	5
10	FAIZAH	1	0	1	1	0	1	0	0	0	1	5
11	FITRI	1	0	1	0	0	1	0	1	0	1	5
12	KYARA	1	0	1	0	0	0	0	1	0	0	3
13	LINGGA	1	0	1	0	0	1	0	1	0	1	5
14	FARHAN	0	1	1	0	0	0	0	0	1	0	3
15	MAHMUDIN	1	1	1	1	1	1	1	0	0	1	8
16	NAYSHA	1	0	1	0	1	1	1	1	0	0	6
17	RAJA	1	0	1	1	1	1	0	0	0	1	6
18	RUSDI	1	0	1	0	1	0	0	1	1	0	5
19	NAJWA	1	1	1	0	0	1	0	1	1	0	6
20	AINI	1	0	1	0	0	1	1	1	0	0	5
	∑x	19	8	20	6	8	13	4	16	4	10	
	∑y											108
	r. hitung	6,051	0,401	0	0,58	0,5	-1,901	0,146	-0,024	-0,22	0,453	
	r. tabel	0,444	0,444	0,444	0,44	0,44	0,444	0,444	0,444	0,444	0,444	
		Valid	Invalid	Invalid	Valid	Valid	Invalid	Invalid	Invalid	Invalid	Valid	

Based on the results of data analysis as shown in the table above, it is known that the r_{count} value for each question ranges from –0.024 to 0.534, while r_{table} = 0.444 at a significance level of 5% with a total of 20 students responding (df = N – 2 = 18). These results indicate that items 1, 2, 3, 4, 6, 8, and 10 have r_{table} values, so they are declared valid, while items 5, 7, and 9 have r_{table} values, so they are categorized as invalid. Thus, it can be concluded that of the ten items tested, seven items were declared valid and three items needed to be revised. This shows that most of the items were able to accurately measure students' abilities in accordance with the learning indicators that had been formulated.

This finding reinforces Creswell's (2018) opinion that validity is related to the level of empirical and theoretical support for test score interpretations. This means that validity is not merely a correlation number, but reflects the extent to which the items actually measure the desired construct. In line with this, Taherdoost (2016) emphasizes that items that have a significant correlation with the total score indicate

consistency between the content of the questions and the abilities being measured. The results of this study are also in line with the findings of Sumintono & Widhiarso (2020), who state that valid instruments can provide accurate and reliable measurement results. Therefore, the seven items that were declared valid can be used in the final test, while the three invalid items need to be revised so that the entire instrument has optimal reliability and validity.

B. Reliability Test Results

Reliability testing is a procedure in research used to measure the extent to which an instrument or measuring tool produces consistent results when used under the same conditions. A reliable instrument indicates that the measurement results are relatively stable, free from random errors, and can be trusted to represent the construct being measured (Livingston et al., 2018). Meanwhile, according to Mueller & Knapp (2018), reliability is not only a characteristic of the test itself, but also a function of the population, context, and data collection methods. Therefore, reliability must be tested empirically in a specific research context.

In the context of social and educational research methodology, Quintão et al. (2020) added that reliability can be improved through systematic research design, assessor training, and the use of instruments that have undergone validation testing. This is important to ensure that the data produced is truly reliable for drawing valid scientific conclusions. In this study, the researchers calculated the reliability test using the following formula:

Description:

$$S^2 = \frac{n\sum x_1^2 - (\sum x_1)^2}{n\,(n-1)} \qquad R_{11} = \left[\frac{m}{m-1}\right].\left[1 - \frac{S^2x}{S^2x}\right]$$

S^2= sample variance
x_i = i-th data value
n = amount of data in the sample
$\sum x_i$ = total amount of data
$\sum x_i^2$ = sum of squares of each data
R_{11} = Test reliability coefficient (level of stability or consistency of the test)
m = Number of questions or items in the test
S_x^2 = Variance of scores for each item (variation between items)
S_r^2 = Total test score variance (overall variation in test results)

In this study, the reliability test was conducted using Microsoft Excel, with data collected from grade XI students at MAN 1 Mandailing Natal in the 2025/2026 academic year, as shown in Table 2.

Tabel 2 . Results of Reliability Testing for Test Instruments

No	Student Name	Question Iteml (X)				Total Valid Score (Y)
		1	4	5	10	
1	ADE	1	1	1	1	4
2	ADIZ	1	0	0	1	2
3	FAJRI	1	0	0	0	1
4	AJIRNI	1	1	1	1	4
5	ALIFAH	1	0	0	0	1

6	ALYA	1	1	1	1	4
7	ANNISA	1	0	0	1	2
8	DIAH	1	0	1	0	2
9	EKA RAMDANI	1	0	0	0	1
10	FAIZAH	1	1	0	1	3
11	FITRI	1	0	0	1	2
12	KYARA	1	0	0	0	1
13	LINGGA	1	0	0	1	2
14	FARHAN	0	0	0	0	0
15	MAHMUDIN	1	1	1	1	4
16	NAYSHA	1	0	1	0	2
17	RAJA	1	1	1	1	4
18	RUSDI	1	0	1	0	2
19	NAJWA	1	0	0	0	1
20	AINI	1	0	0	0	1
	∑	19	6	8	10	43
	S^2	0,05	0,221	0,253	0,263	
	S^2T					0,6
	S^2 X total					2,567
	R_11	1,307	1,218	1,202	1,197	

Based on the results of data analysis in Table 2 of the reliability test above, the R_{11} value or reliability coefficient obtained for the three items (numbers 4, 5, and 10) are 1.218, 1.202, and 1.197, respectively. These values indicate a very high level of consistency, even exceeding the ideal reliability limit (generally between 0.70 and 0.90). This indicates that the instrument used is very consistent in measuring students' abilities in relation to the material being tested. However, the R_{11} value above 1 may also be due to rounding or technical errors in the calculation of item variance, because theoretically the reliability value should not exceed 1 (George & Mallery, 2018). Nevertheless, in general, the data show that the items have a strong level of stability among students, so they can be declared reliable as a learning measurement tool.

In the context of educational research, high reliability indicates that the instrument can produce consistent scores when administered to similar groups of students under the same conditions (Plonsky & Derrick, 2016). Meanwhile, according to Quaigrain & Arhin (2017), good reliability indicates that each item contributes significantly to the overall measurement, so that the instrument can be trusted for learning evaluation. Therefore, from the results of the table analysis, it can be concluded that the test instrument in this study meets the criteria for excellent reliability and is suitable for use in academic assessment.

C. Results of the Question Difficulty Level Analysis

Difficulty level testing is one type of item analysis conducted with the aim of determining the level of difficulty of a test in measuring student ability. The difficulty level describes the proportion of students who can answer a question correctly, making it an important indicator in determining the quality of a test instrument. According to Mudiyntri and Indriani (2025), the difficulty level is the probability of a student answering an item correctly, which indicates the extent to which the item is easy or

difficult. An item is considered acceptable if its difficulty level is not too difficult or too easy. This means that a proportional difficulty level indicates a quality item. The formula used to measure the difficulty level is:

Description:

$$P = \frac{\bar{S}}{S_{maks}}$$

P : Difficulty index
S : Average score on each item
Smaks : Maximum score for that item

The results of the analysis of the difficulty level of the questions in the test instrument for the subject of Al-Quran Hadith with the material of Da'wah, Khutbah, and Tablig that was given to 11th grade students at MAN 1 Mandailing Natal in the 2025/2026 academic year. This analysis was carried out using Microsoft Excel, which is shown in Table 3.

Table 3. Results of item difficulty level analysis

Question Number	Many students answered correctly (NP)	Question difficulty index (P) = $\frac{NP}{N}$	Description
1	19	0.95	Easy
2	8	0.40	Currently
3	20	1.00	Easy
4	6	0.30	Currently
5	8	0.40	Currently
6	13	0.65	Currently
7	4	0.20	Difficult
8	16	0.80	Easy
9	4	0.20	Sukar
10	10	0.50	Currently

Based on the results of the difficulty level analysis in the table, it was found that 30% of the questions were easy, 50% were medium, and 20% were difficult. These proportions show that most of the questions were of medium difficulty, so that in general the quality of the questions can be categorized as good. Questions with a moderate level of difficulty are considered ideal because they are able to provide an accurate picture of students' abilities. They are not too easy, so that all students answer correctly, nor are they too difficult, so that many students fail to answer. This proportion also supports the principle of learning evaluation that requires a balance between discriminating power and level of difficulty, so that the assessment instrument is able to measure students' abilities optimally (Hanum et al., 2025).

In addition, the inclusion of easy and difficult items is also necessary to maintain a variety of difficulty levels, which can help teachers differentiate between the abilities of students at various cognitive levels. However, items that are too easy (P > 0.90), such as number 3, and those that are too difficult (P < 0.25), such as number 9, need to be reviewed to avoid bias in the assessment results. Thus, it can be concluded that this test instrument has a good balance between easy, medium, and difficult items, which means that the test is of sufficient quality to measure student learning outcomes in a

representative manner

D. Results of Discrimination Analysis

Discrimination index is one of the main components in item analysis used to determine the extent to which an item can distinguish between high and low ability test takers. In other words, this test measures how well an item can identify differences in material mastery among students. According to Suryani and Nasir (2024), discrimination indicates the ability of an item to separate groups of high- and low-ability participants based on the proportion of correct answers. Items with high discrimination will be answered correctly by high-ability participants and incorrectly by low-ability participants. The following formula is used to analyze discrimination:

Description:

$$D = \frac{BA - BB}{J}$$

BA : Number of correct answers in the top group
BB: Number of correct answers in the lower group
J : Number of students in each group

The results of the analysis of the discriminating power of the test items in the test instrument for the subject of Al-Quran Hadith with the material of Da'wah, Khutbah, and Tablig that had been given to 11th grade students at MAN 1 Mandailing Natal in the 2025/2026 academic year. This analysis was carried out using Microsoft Excel, which is shown in Table 4.

Tabel 4. Results of discriminant analysis on test items

Question Number	BA	BB	D	Category
1	6	5	0.17	Bad
2	5	2	0.5	Good
3	6	6	0.0	Bad
4	5	1	0.67	Good
5	5	1	0.67	Good
6	6	2	0.67	Good
7	3	1	0.33	Enough
8	6	3	0.5	Good
9	2	1	0.17	Bad
10	5	2	0.5	Good

Based on the results of the discriminating power analysis in the table above, it is known that out of ten questions, there are six questions (numbers 2, 4, 5, 6, 8, and 10) that are in the good category, one question (number 7) in the fair category, and three questions (numbers 1, 3, and 9) in the poor category. In general, most of the items have good discriminating power, indicating that the items are able to effectively distinguish between high and low ability students. The discrimination values (D) ranging from 0.33 to 0.67 in most items indicate that this test instrument has good reliability potential. This finding is in line with the results of research by Susanti and Darmawan (2021), which states that items with high discrimination values contribute greatly to the validity and quality of learning evaluation tools.

However, there were three items (numbers 1, 3, and 9) with a poor category (D < 0.20), indicating that these items were not able to optimally distinguish between high- and low-ability students. Therefore, these items should be revised or replaced with questions that are more representative of the competencies being measured. Overall, the results of the discrimination analysis show that this test instrument is of fairly good quality, as most of the items have good discrimination and only a few need improvement. This is in line with the opinion of Fitriani et al. (2019) that the composition of items with high discrimination is an important indicator in maintaining the balance between the reliability and validity of learning outcome test instruments.

CONCLUSION AND SUGGETIONS

Based on the results of the analysis of the validity, reliability, difficulty level, and discriminating power of multiple-choice questions on the subject matter of Da'wah, Khutbah, and Tablig for grade XI at MAN 1 Mandailing Natal, it can be concluded that, in general, the test instruments are of sufficient quality and are suitable for use as learning evaluation tools. The validity test results show that of the ten questions analyzed, seven questions were declared valid and three questions needed to be revised because they had a low correlation with the total score. This means that most of the questions were able to accurately measure the expected competencies.

Furthermore, the reliability test results show that the reliability coefficient (R_{11}) obtained is very high, indicating that the test instrument is consistent and stable when used to measure students' abilities in the same context. Although there is a possibility of rounding errors in the calculation results, the data still shows an excellent level. The analysis of the level of difficulty of the questions shows that 30% of the questions are in the easy category, 50% are medium, and 20% are difficult. This composition shows an ideal proportion because a balanced distribution of questions allows for a more representative measurement of student abilities. The results of the discrimination power analysis show that most of the questions (60%) are in the good category, one question is adequate, and three questions need to be revised due to low discrimination power. Thus, most of the questions have been able to effectively distinguish between high and low ability students.

Overall, it can be concluded that the multiple-choice questions on the topics of Da'wah, Khutbah, and Tablig for grade XI at MAN 1 Mandailing Natal have met the criteria of validity, reliability, level of difficulty, and good discriminating power. However, to improve the quality of the evaluation, questions that are invalid and have low discriminating power need to be revised or replaced. This instrument is suitable for use as an objective, reliable, and representative assessment tool for students' spiritual abilities.

REFERENCES

Alfani, N. (2022). Analysis of Summative Tests in Secondary School Mathematics Learning. Journal of Educational Evaluation, 14(2), 45–58.

Anderson, L. W. (2019). Assessment in Education: Principles and Applications. New York: Routledge.

Arikunto, S. (2018). Research Procedures: A Practical Approach (7th Revised Edition, pp. 120–121). Jakarta: Rineka Cipta.

Arikunto, S. (2018). Research Procedures: A Practical Approach (Revised Edition, pp. 120–121). Jakarta: Rineka Cipta.

Astuti, R. (2018). Analysis of Multiple Choice Items in Learning Evaluation. Journal of Educational Research, 35(1), 23–31.

Creswell, J. W., & Creswell, J. D. (2017). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (4th ed.). Thousand Oaks, CA: Sage Publications. ISBN: 978-1-5063-8670-6.

Crocker, L., & Algina, J. (2015). Introduction to Classical and Modern Test Theory. New York: Cengage Learning.

Fauziyyah, N. (2019). The relationship between measurement, assessment, and evaluation in education: A theoretical study. Scientific Journal of Education and Learning, 3(1), 45–53.

Fitriani, N., Sari, L., & Prasetyo, A. (2019). Analysis of final semester exam questions for Islamic Religious Education in senior high school. Journal of Educational Evaluation, 10(1), 45–56.

George, D., & Mallery, P. (2018). Reliability analysis. In IBM SPSS Statistics 26 step by step: A simple guide and reference (pp. 241–255). Routledge. https://doi.org/10.4324/9781351033909-25

George, D., & Mallery, P. (2018). SPSS for Windows Step by Step: A

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and Validating Test Items. New York: Routledge.

Hanum, F., Muliono, T. P. R. Z., & Safira, F. (2025). Development of a test instrument for creative thinking skills in preaching material for high school students. Relativitas Journal: Innovation in Aqidah Learning, 7(1), 45–54. Malikussaleh University.

Hayati, S. (2016). The influence of instrument validity and reliability on educational research results. Journal of Educational Research and Evaluation, 20(1), 55–63.

Ida, I., & Musyarofah, M. (2021). Analysis of the quality of multiple-choice questions as an effort to improve learning evaluation instruments. Journal of Educational Measurement and Evaluation, 11(1), 34–42.

Livingston, S. A., Carlson, J., & Bridgeman, B. (2018). Test reliability: Basic concepts. ETS Research Memorandum, RM-18-01, 1–12. Educational Testing Service (ETS).

Lubis, W. A. (2024). *Pengembangan E-Komik melalui Flipbuilder untuk Meningkatkan Minat Belajar Peserta Didik pada Mata Pelajaran Sejarah Kebudayaan Islam* [Sekolah Tinggi Agama Islam Negeri Mandailing Natal]. https://repository.stain-madina.ac.id/id/eprint/212

Lubis, W. A. (2026). Pengembangan Media Pembelajaran Gamifikasi Berbasis Quizizz untuk Meningkatkan Motivasi Belajar Peserta Didik. *An-Nuha*, *6*(1), 78–93. https://doi.org/https://doi.org/10.24036/annuha.v6i1.745

Mardapi, D. (2008). Techniques for Developing Test and Non-Test Instruments. Yogyakarta: Mitra Cendekia Press.

Mudiyntri, N. S., & Indriani, N. S. (2025). Testing the quality of financial professional ethics questions as a learning evaluation tool for grade XI through the Anates version 4.0 test. Journal of Economic Education, 10(2).

Mueller, R. O., & Knapp, T. R. (2018). Reliability and validity. In Quantitative research methods for professionals (pp. 423–438). Routledge. https://doi.org/10.4324/9781315755649-29

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

Novia, A., Susanti, R., & Wibowo, D. (2020). Analysis of Multiple Choice Question Quality in Terms of Validity and Reliability. Journal of Learning Evaluation, 8(2), 112–121.

Novia, R., Putri, A. N., & Fitria, D. (2020). Validity and reliability of multiple-choice questions in religious education. Journal of Islamic Education and Teaching, 8(2), 120–130.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. Modern Language Journal, 100(3), 713–739. https://doi.org/10.1111/modl.12335

Popham, W. J. (2017). Classroom Assessment: What Teachers Need to Know (8th ed.). Boston: Pearson.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. Cogent Education, 4(1), 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Quintão, C., Andrade, P., & Almeida, F. (2020). How to improve the validity and reliability of a case study approach? Journal of International Students in Education, 10(3), 45–60.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. Psychological Assessment, 31(12), 1395–1411. https://doi.org/10.1037/pas0000754

Sumintono, B., & Widhiarso, W. (2020). Application of the Rasch model for social science research (revised edition). Trim Komunikata Publishing House.

Suryani, L., & Nasir, A. (2024). Analysis of Islamic Education learning outcome test items based on discriminating power and difficulty level. Journal of Islamic Education

Taherdoost, H. (2016). Validitas dan reliabilitas alat penelitian: Cara menguji validitas kuesioner/survei dalam penelitian. Jurnal Penelitian Akademik dalam Manajemen, 5(3), 28–36.