

The effectiveness of the *tahriri* examination as an instrument for measuring students' understanding at Al-Hikmah Tulungagung Islamic Boarding School

Eva Kumalasari^{1*}, Nayla Wirdatun Ni'ammah², Chusnul Chotimah³

^{1 2 3} UIN Sayyid Ali Rahmatullah, Tulungagung, Indonesia

*Corresponding Author: esari9938@gmail.com

Submission: 12 June 2026

Revision: 2 July 2026

Accepted: 13 July 2026

Abstract

The effectiveness of the *tahriri* examination as an assessment instrument in Islamic boarding schools remains debated because empirical evidence linking it to students' level of understanding is still limited. This study aimed to describe students' level of understanding, evaluate the effectiveness of the *tahriri* examination, and analyze the relationship between the *tahriri* examination and students' level of understanding at Al-Hikmah Tulungagung Islamic Boarding School for University Students. A quantitative descriptive-correlational design was employed involving 57 *santri* selected through purposive sampling, with data analyzed using descriptive statistics and Spearman's rank correlation. The findings revealed that the instrument met the validity requirements (9 of 10 items were valid) and demonstrated high reliability (*Cronbach's Alpha* = 0.866). The effectiveness of the *tahriri* examination reached 87.6%, students' level of understanding reached 86.9%, and a strong positive and statistically significant relationship was identified between the two variables ($r = 0.757$, $p < 0.001$). This study provides empirical evidence supporting the effectiveness of the *tahriri* examination and proposes a conceptual framework for improving learning assessment in Islamic boarding schools.

Keywords: *Tahriri* examination; Learning assessment; Students' level of understanding; *Santri*; Islamic boarding school

Abstrak

Ketepatan ujian *tahriri* sebagai instrumen evaluasi pembelajaran di pesantren masih diperdebatkan karena belum banyak penelitian yang membuktikan hubungannya dengan tingkat pemahaman santri secara empiris. Penelitian ini bertujuan mendeskripsikan tingkat pemahaman santri, mengukur efektivitas ujian *tahriri*, serta menganalisis hubungan antara ujian *tahriri* dan tingkat pemahaman santri di Pesantren Mahasiswa Al-Hikmah Tulungagung. Penelitian menggunakan pendekatan kuantitatif dengan desain deskriptif-korelasional terhadap 57 santri yang dipilih melalui *purposive sampling*, kemudian dianalisis menggunakan statistik deskriptif dan korelasi Spearman Rank. Hasil penelitian menunjukkan bahwa instrumen memenuhi kriteria validitas (9 dari 10 butir valid) dan reliabilitas tinggi (*Cronbach's Alpha* = 0,866), tingkat efektivitas ujian *tahriri* mencapai 87,6%, tingkat pemahaman santri sebesar 86,9%, serta terdapat hubungan positif yang kuat dan signifikan antara kedua variabel ($r = 0,757$; $p < 0,001$). Penelitian ini memberikan bukti empiris bahwa ujian *tahriri* merupakan instrumen evaluasi yang efektif serta menawarkan kerangka konseptual sebagai acuan pengembangan evaluasi pembelajaran di pesantren.

Kata Kunci: Ujian *tahriri*; Evaluasi pembelajaran; Tingkat pemahaman; Santri; Pesantren.

Introduction

The accuracy of assessment instruments remains one of the major challenges in the learning process because assessment results are frequently used as the basis for determining students' level of subject mastery (Taber, 2018). The problem arises when assessment instruments fail to comprehensively capture students' level of understanding, meaning that the scores obtained do not necessarily reflect their actual competencies (Bao et al., 2022). This issue is also evident in Islamic boarding schools, where the *tahriri* examination continues to serve as the primary assessment method for measuring students' learning achievement (Awang et al., 2025). On the one hand, the *tahriri* examination is considered effective because it is easy to administer, follows a systematic scoring procedure, and enables educators to assess a large number of students within a relatively short period (Febriani & Widayanti, 2021). On the other hand, written examinations are often criticized for emphasizing memorization rather than assessing deeper conceptual understanding, making their effectiveness as assessment instruments an issue that still requires empirical verification.

This issue is particularly significant because the quality of assessment directly influences the quality of academic decisions made by educators. Assessment instruments that fail to measure students' understanding accurately may produce biased information, thereby affecting instructional improvement, the development of teaching strategies, and the evaluation of students' learning achievement (Kaur et al., 2020). Within the context of Islamic boarding schools, the implications are even more substantial because learning success is evaluated not only through students' mastery of Islamic knowledge but also through their ability to understand, explain, and apply the concepts they have learned. Therefore, the effectiveness of the *tahriri* examination should not be judged solely by its practicality or ease of implementation but rather by the extent to which it objectively, accurately, and consistently represents students' level of understanding.

Al-Hikmah Tulungagung Islamic Boarding School for University Students employs the *tahriri* examination as one of its primary instruments for evaluating students' learning outcomes at the end of each instructional period. The assessment process begins with the development of examination items based on the instructional content, after which the examination results are used as indicators of learning achievement. In practice, this process encompasses test development, examination administration, score evaluation, and the assessment of students' learning outcomes. To ensure that assessment results accurately reflect students' level of understanding, the instrument must satisfy the principles of validity, reliability, and objectivity while being capable of measuring students' abilities to remember, understand, and analyze the learning materials proportionately. Consequently, the effectiveness of the *tahriri* examination

depends not only on the administration of the test itself but also on the overall quality of assessment design and implementation throughout the learning process.

Research on learning assessment through written examinations has been extensively conducted across various educational levels. Research by Sagala and Andriani (2019) demonstrated that written examinations effectively measure learning achievement when they are developed based on clearly defined learning objectives and assessment indicators. Likewise, research by Alfiyah et al. (2025) found that instrument quality is the primary factor determining the accuracy of assessment results, emphasizing that validity and reliability must be established before an instrument is implemented. Furthermore, research by Halperin (2020) revealed a significant relationship between written examination scores and students' mastery of learning materials; however, the study focused primarily on formal schools and higher education institutions. Meanwhile, research by Seyaningsih and Sari (2021) highlighted the importance of developing assessment instruments capable of measuring higher-order thinking skills but did not specifically examine the effectiveness of the *tahriri* examination within the context of Islamic boarding schools. Therefore, a research gap remains regarding how the quality of *tahriri* examination instruments relates to students' level of understanding in Islamic boarding schools, which possess educational characteristics distinct from those of formal educational institutions.

Based on these considerations, this study aims to describe students' level of understanding based on the results of the *tahriri* examination, evaluate the effectiveness of the *tahriri* examination as a learning assessment instrument, and analyze the relationship between the *tahriri* examination and students' level of understanding at Al-Hikmah Tulungagung Islamic Boarding School for University Students. This study is significant because it provides empirical evidence regarding the effectiveness of the *tahriri* examination, which has long served as a fundamental component of the assessment system in Islamic boarding schools. Unlike previous studies that primarily focused either on learning outcomes or instrument quality independently, this study integrates instrument quality testing, statistical relationship analysis, and the development of a conceptual framework explaining the effectiveness of the *tahriri* examination as an instrument for measuring students' level of understanding. Consequently, the findings are expected not only to enrich the literature on learning assessment but also to provide a foundation for developing more accurate, objective, and contextually appropriate assessment systems for Islamic boarding schools.

Method

This study employed a quantitative approach using a descriptive-correlational design to describe students' level of understanding, evaluate the effectiveness of the *tahriri* examination as a learning assessment instrument, and analyze the relationship

between the *tahriri* examination and students' level of understanding at Al-Hikmah Tulungagung Islamic Boarding School for University Students (Creswell, 2014). The independent variable was the *Tahriri* Examination (X), whereas the dependent variable was Students' Level of Understanding (Y). The study population comprised all *santri* who actively participated in the learning process and had taken the *tahriri* examination. A total of 57 *santri* were selected through purposive sampling based on the following criteria: having taken the *tahriri* examination at least once during the most recent academic period, voluntarily agreeing to participate in the study, and possessing complete examination score records (Hair et al., 2021). Data were collected through documentation of *tahriri* examination scores and a 10-item Likert-scale questionnaire designed to measure perceptions of the quality of the *tahriri* examination, including the alignment of test items with learning objectives, item clarity, assessment objectivity, and the ability of the examination to measure students' level of understanding.

The research procedure consisted of four stages: preparation, data collection, data analysis, and conclusion drawing. During the preparation stage, the research instrument was developed and subsequently evaluated using the Pearson Product-Moment validity test and Cronbach's Alpha reliability test. The collected data were then analyzed using IBM SPSS Statistics. Descriptive statistics were employed to describe the characteristics of each research variable, while the Kolmogorov-Smirnov test was conducted to examine data normality before the correlation analysis. Since the normality test indicated that the data were not normally distributed, the relationship between the *Tahriri* Examination and Students' Level of Understanding was analyzed using Spearman's rank correlation at a significance level of 5% ($\alpha = 0.05$). This statistical approach was selected because it was appropriate for the distributional characteristics of the data and enabled an objective assessment of both the direction and strength of the relationship between the two variables (Ghozali, 2006).

Results and Discussion

Instrument testing and analysis prerequisite results

Before examining the relationship between the *Tahriri* Examination and Students' Level of Understanding, the research instrument and the analytical assumptions were evaluated to ensure that the collected data met the required statistical criteria. Instrument testing was conducted to determine the validity and reliability of each questionnaire item, thereby ensuring that the resulting data could serve as a trustworthy basis for drawing research conclusions. In addition, descriptive statistical analysis was employed to describe the characteristics of the respondents' data, while a normality test was performed to determine the most appropriate inferential statistical technique. The overall results demonstrated that the research instrument satisfied the requirements of a valid and reliable measurement tool for investigating

the effectiveness of the *tahriri* examination in measuring students' level of understanding. The results of the instrument validity test are presented in Table 1.

Table 1. Instrument validity test results

Item	Correlation Coefficient (r)	Sig.	Interpretation
P01-P07, P09-P10	0.325-0.869	<0.05	Valid
P08	-0.039	0.775	Invalid

Source: Processed by the authors (2026).

As presented in Table 1, nine of the ten questionnaire items obtained correlation coefficients ranging from 0.325 to 0.869 with significance values below 0.05, thereby satisfying the validity criteria. In contrast, item P08 yielded a negative correlation coefficient of -0.039 with a significance value of 0.775 and was therefore classified as invalid and excluded from subsequent analyses. These findings indicate that most of the indicators adequately represented the research construct and could be used to measure the variables under investigation. Since the validity requirements were fulfilled, the research instrument was considered capable of accurately representing the respondents' empirical conditions in accordance with the objectives of the study.

After establishing instrument validity, the next step involved reliability testing to determine the internal consistency of the instrument when applied under relatively similar conditions. Reliability is an essential indicator because it reflects the extent to which an instrument produces stable and dependable measurements. The higher the reliability coefficient, the lower the likelihood of inconsistencies in the measurement process. Therefore, reliability was assessed using Cronbach's Alpha, one of the most widely accepted methods in quantitative research. The results of the reliability analysis are presented in Table 2.

Table 2. Instrument reliability test results

Instrument	Cronbach's Alpha	Interpretation
Research questionnaire	0.866	Highly reliable

Source: Processed by the authors (2026).

Referring to Table 2, the research instrument obtained a Cronbach's Alpha coefficient of 0.866, indicating excellent reliability because it exceeded the recommended threshold of 0.70. This result demonstrates that all valid questionnaire items possessed a high level of internal consistency and were capable of producing relatively stable measurements when administered to respondents with similar characteristics. Accordingly, the instrument fulfilled the reliability requirements and was considered appropriate for data collection. Once the quality of the instrument had been confirmed, the analysis proceeded to descriptive statistics and normality testing to describe the distribution of the data and determine the most appropriate inferential statistical

technique. A summary of the descriptive statistics is presented in Table 3, while the results of the normality test are shown in Table 4.

Table 3. Descriptive statistics

Variable	Mean	Minimum	Maximum	Standard Deviation
<i>Tahriri</i> Examination (X)	26.07	14	30	3.760
Students' Level of Understanding (Y)	14.02	9	16	1.706

Source: Processed by the authors (2026).

As shown in Table 3, the *Tahriri* Examination variable obtained a mean score of 26.07 with a standard deviation of 3.760, whereas Students' Level of Understanding achieved a mean score of 14.02 with a standard deviation of 1.706. These mean values indicate that both variables were generally within the high category, suggesting that the respondents demonstrated favorable perceptions of the *tahriri* examination while also exhibiting a relatively high level of understanding. Furthermore, the relatively small standard deviation values indicate that the data were fairly homogeneous, reflecting a consistent distribution of responses across the study participants.

Table 4. Normality test results

Method	Sig.	Interpretation
Kolmogorov-Smirnov	0.004	Not normally distributed

Source: Processed by the authors (2026).

As presented in Table 4, the Kolmogorov-Smirnov test produced a significance value of 0.004, which was lower than the established significance level of 0.05. This result indicates that the data were not normally distributed. Consequently, the relationship between the variables was analyzed using Spearman's rank correlation, a non-parametric statistical technique appropriate for the distributional characteristics of the data. Therefore, the instrument testing and prerequisite analyses provided a sound methodological foundation for examining the relationship between the *tahriri* examination and students' level of understanding.

Relationship between the tahriri examination and students' level of understanding

After the research instrument had been confirmed to be valid and reliable and the data had satisfied the analytical prerequisites, the next stage was to examine the relationship between the *Tahriri* Examination and Students' Level of Understanding. This analysis aimed to determine whether a significant relationship existed between the two research variables among *santri* at Al-Hikmah Tulungagung Islamic Boarding School for University Students. Based on the results of the normality test presented in the previous subsection, the data were not normally distributed; therefore, the relationship between the variables was analyzed using Spearman's rank correlation as a non-parametric statistical technique. This approach enabled the researchers to

identify the direction, strength, and statistical significance of the relationship between the variables. A summary of the Spearman's rank correlation analysis is presented in Table 5.

Table 5. Spearman's rank correlation results

Variables	Correlation Coefficient (r)	Sig. (2-tailed)	Interpretation
<i>Tahriri</i> Examination ↔ Students' Level of Understanding	0.757	<0.001	Strong positive relationship

Source: Processed by the authors (2026).

As presented in Table 5, Spearman's rank correlation analysis yielded a correlation coefficient of 0.757 with a significance value of <0.001. These findings indicate a strong positive relationship between the *tahriri* examination and students' level of understanding. The positive direction of the relationship suggests that improvements in the quality of the *tahriri* examination tend to be accompanied by higher levels of students' understanding of the learning materials. Moreover, the significance value, which was lower than the established significance level of 0.05, confirms that the observed relationship is statistically significant and unlikely to have occurred by chance. Accordingly, the results provide empirical evidence that the two variables are strongly associated within the learning context of Al-Hikmah Tulungagung Islamic Boarding School for University Students.

To provide a clearer interpretation of the magnitude of the relationship between the variables, the correlation coefficient was further interpreted using the coefficient of determination (r^2). The calculation produced a value of 57.3%, indicating that 57.3% of the variation in the relationship between the *tahriri* examination and students' level of understanding was associated with the variables examined in this study, while the remaining 42.7% was related to other factors not included in the analysis. These findings suggest that the relationship between the two variables can be classified as strong, although students' level of understanding is also influenced by other factors, such as learning motivation, instructional strategies employed by teachers, the learning environment, and individual student characteristics. Therefore, the correlation results should be interpreted as evidence of the strength of the association between the variables rather than as proof of a causal relationship.

Overall, the correlation analysis demonstrates a strong and statistically significant positive relationship between the *tahriri* examination and students' level of understanding. These findings indicate that the quality of the *tahriri* examination is closely associated with students' ability to understand the learning materials. Nevertheless, because this study employed a correlational design, the findings only demonstrate the existence of a statistical relationship between the variables without explaining the underlying mechanisms or causal factors responsible for that

relationship. Therefore, the implications of these findings and their relevance to the role of the *tahriri* examination as a learning assessment instrument are discussed in greater depth in the following subsection.

Effectiveness of the tahriri examination as an instrument for measuring students' level of understanding

One of the ongoing debates in learning assessment concerns whether written examinations are still capable of comprehensively measuring students' level of understanding or merely reflecting their ability to memorize information. Muslimah et al. (2025) argued that the *tahriri* examination tends to assess students' ability to reproduce information and therefore does not fully capture their conceptual mastery. In contrast, Schuwirth and van der Vleuten (2018) maintained that written examinations remain relevant when they are designed based on well-defined learning objectives, clear assessment indicators, and established principles of validity and reliability. These contrasting perspectives suggest that the effectiveness of the *tahriri* examination cannot be evaluated solely based on its assessment format but must also consider the quality of the assessment instrument itself. Based on the findings of the present study, the relationships among these components can be understood through the conceptual framework presented in Figure 1.

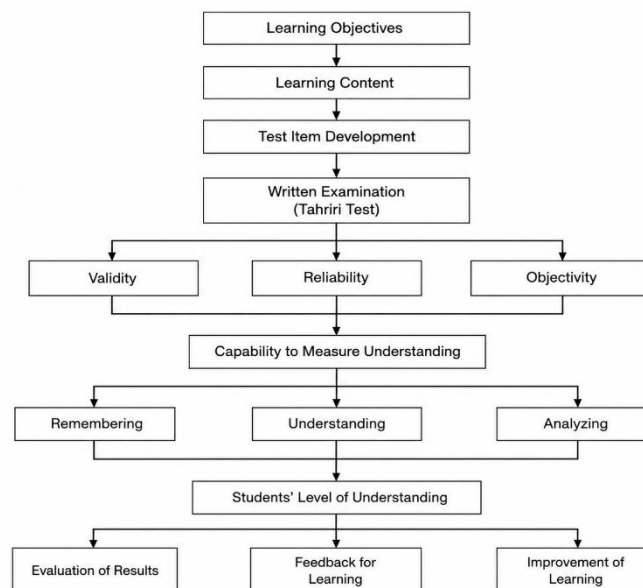


Figure 1. Conceptual framework of the effectiveness of the *tahriri* examination in measuring students' level of understanding

Source: Developed by the authors based on the research findings (2026).

Figure 1 illustrates that the effectiveness of the *tahriri* examination results from the interrelationship among learning objectives, assessment instrument development, measurement quality, and the examination's ability to evaluate students' cognitive processes. The framework demonstrates that test items developed in alignment with learning objectives produce assessment instruments that satisfy the principles of

validity, reliability, and objectivity, thereby enabling a more accurate assessment of students' abilities to remember, understand, and analyze learning materials. Consequently, the effectiveness of the *tahriri* examination should not be interpreted merely through examination scores but rather as a comprehensive assessment process that connects instrument quality with students' level of understanding. The conceptual framework also explains why the relationship identified in this study exhibited a strong correlation.

The first aspect illustrated in Figure 1 is the relationship among *learning objectives*, *learning content*, and *test item development*. These three components indicate that assessment effectiveness largely depends on the alignment between instructional objectives, the content delivered, and the indicators measured through examination items. Consistent with the findings of Yang and Li (2018), assessment instruments developed according to intended learning outcomes provide a more accurate measurement of students' competencies. Likewise, Khan et al. (2025) explained that misalignment between learning objectives and assessment items reduces the ability of examinations to accurately represent students' actual competencies. The findings of the present study reinforce those reported by Aflalo (2021), demonstrating that high-quality test construction constitutes the fundamental basis upon which the effectiveness of examinations can be evaluated through students' learning outcomes.

The next components of the framework are *validity*, *reliability*, and *objectivity*, which represent the essential indicators of assessment instrument quality. These components demonstrate that an effective assessment instrument should not only measure the intended construct accurately but also produce consistent and objective measurement results. Consistent with the findings of Ma et al. (2020), validity and reliability are indispensable prerequisites for developing trustworthy assessment instruments. Similarly, Immonen et al. (2019) emphasized that assessment objectivity plays a crucial role in enhancing the credibility of evaluation results. Unlike Hughes (2018), who identified validity as the dominant indicator of assessment quality, the present study demonstrates that validity, reliability, and objectivity function as complementary dimensions that collectively determine the overall quality of an assessment instrument. This conclusion is further supported by the empirical findings, which confirmed that the research instrument achieved satisfactory validity and high reliability before the relationship between the variables was examined.

Figure 1 further demonstrates that the *tahriri* examination measures not only *remembering* but also *understanding* and *analyzing* as essential dimensions of students' cognitive processes. These findings indicate that the effectiveness of the *tahriri* examination depends on the ability of the assessment items to evaluate conceptual understanding rather than merely measuring factual recall. Consistent with the findings of Yu (2023), effective learning assessment should encompass multiple levels

of cognitive ability to ensure that learning outcomes accurately reflect students' mastery of the subject matter. Likewise, Heber et al. (2023) found that examinations primarily emphasizing memorization exhibit limited discriminative power in distinguishing students' levels of understanding. The findings of the present study strengthen this argument, as the strong positive relationship identified between the *tahriri* examination and students' level of understanding indicates that well-designed assessment instruments are capable of representing more complex cognitive competencies.

Overall, this study demonstrates that the effectiveness of the *tahriri* examination as an instrument for measuring students' level of understanding is supported not only by a strong statistical relationship but also by the integration of learning objectives, high-quality test development, sound assessment instrument quality, and the examination's ability to assess students' cognitive processes, as illustrated in Figure 1. These findings indicate that the *tahriri* examination functions not only as a tool for determining students' mastery of learning materials but also as a valuable source of information for evaluating learning outcomes and improving instructional practices. Therefore, the effectiveness of the *tahriri* examination lies not merely in its assessment procedure but in its capacity to generate accurate, objective, and meaningful information that supports evidence-based decision-making within the context of Islamic boarding school education.

Conclusion

This study demonstrates that the *tahriri* examination is an effective assessment instrument for measuring students' level of understanding at Al-Hikmah Tulungagung Islamic Boarding School for University Students. This conclusion is supported by the research instrument, which satisfied the validity criteria with 9 out of 10 questionnaire items declared valid and exhibited excellent reliability (Cronbach's Alpha = 0.866). Furthermore, the effectiveness of the *tahriri* examination reached 87.6%, while students' level of understanding reached 86.9%, both of which were classified as very high. Spearman's rank correlation analysis revealed a strong and statistically significant positive relationship between the *tahriri* examination and students' level of understanding ($r = 0.757$, $p < 0.001$). These findings indicate that the effectiveness of the *tahriri* examination is determined not only by the implementation of the assessment itself but also by the quality of instrument development, its alignment with learning objectives, and its ability to comprehensively measure students' conceptual understanding.

This study contributes empirical evidence to the development of learning assessment systems in Islamic boarding schools by reinforcing the view that the *tahriri* examination remains a relevant instrument for measuring students' level of understanding when it is developed according to the principles of validity, reliability,

and objectivity. In addition, this study proposes a conceptual framework that explains the relationship between assessment instrument quality and students' level of understanding, providing a practical reference for improving learning assessment practices in Islamic boarding schools. Nevertheless, this study is limited to a single research site involving 57 respondents and focuses solely on the relationship between the *tahriri* examination and students' level of understanding without considering other factors that may influence learning outcomes. Therefore, future studies are recommended to include larger sample sizes, investigate different types of Islamic boarding schools, and incorporate additional variables, such as learning motivation, instructional strategies, and higher-order thinking skills, to provide a more comprehensive understanding of the effectiveness of learning assessment.

References

- Aflalo, E. (2021). Students generating questions as a way of learning. *Active Learning in Higher Education*, 22(1), 63–75. <https://doi.org/10.1177/1469787418769120>
- Alfiyah, N., Zuhriyah, I. A., & Chalishah, N. D. (2025). Validity and reliability analysis: Measuring the accuracy and consistency of instruments for the evaluation of Islamic religious education tests. *JlIP - Jurnal Ilmiah Ilmu Pendidikan*, 8(3), 2672–2679. <https://doi.org/10.54371/jiip.v8i3.7230>
- Awang, A. B., Jalil, M. S. B., Satari, H. B., & Ahmad, M. H. S. B. (2025). The effectiveness of the Plus Tahfiz programme in strengthening Quran memorisation and its impact on students' academic performance. *Al-Bayan: Journal of Qur'an and Hadith Studies*, 23(3), 426–446. <https://doi.org/10.1163/22321969-12342304>
- Bao, L., Koenig, K., Xiao, Y., Fritchman, J., Zhou, S., & Chen, C. (2022). Theoretical model and quantitative assessment of scientific thinking and reasoning. *Physical Review Physics Education Research*, 18(1), 010115. <https://doi.org/10.1103/PhysRevPhysEducRes.18.010115>
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage.
- Febriani, S. R., & Widayanti, R. (2021). Evaluation of Arabic learning based on multiple intelligences classroom. *International Journal of Arabic Language Teaching*, 2(02), 131. <https://doi.org/10.32332/ijalt.v2i02.2486>
- Ghozali, I. (2006). *Aplikasi analisis multivariate dengan program SPSS*. Badan Penerbit Universitas Diponegoro.
- Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). *Partial least squares structural equation modeling (PLS-SEM) using R*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-80519-7>

- Halperin, A. (2020). Mastery-based testing in calculus with a final exam component. *PRIMUS*, 30(8–10), 1017–1039. <https://doi.org/10.1080/10511970.2019.1700575>
- Heber, S., Grasl, M. C., & Volf, I. (2023). A successful intervention to improve conceptual knowledge of medical students who rely on memorization of disclosed items. *Frontiers in Physiology*, 14. <https://doi.org/10.3389/fphys.2023.1258149>
- Hughes, D. J. (2018). Psychometric validity. In *The Wiley handbook of psychometric testing* (pp. 751–779). Wiley. <https://doi.org/10.1002/9781118489772.ch24>
- Immonen, K., Oikarainen, A., Tomietto, M., Kääriäinen, M., Tuomikoski, A.-M., Kaučič, B. M., Filej, B., Riklikiene, O., Flores Vizcaya-Moreno, M., Perez-Cañaveras, R. M., De Raeve, P., & Mikkonen, K. (2019). Assessment of nursing students' competence in clinical practice: A systematic review of reviews. *International Journal of Nursing Studies*, 100, 103414. <https://doi.org/10.1016/j.ijnurstu.2019.103414>
- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3313831.3376219>
- Khan, H. F., Qayyum, S., Beenish, H., Khan, R. A., Iltaf, S., & Faysal, L. R. (2025). Determining the alignment of assessment items with curriculum goals through document analysis by addressing identified item flaws. *BMC Medical Education*, 25(1), 200. <https://doi.org/10.1186/s12909-025-06736-4>
- Ma, L.-L., Wang, Y.-Y., Yang, Z.-H., Huang, D., Weng, H., & Zeng, X.-T. (2020). Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: What are they and which is better? *Military Medical Research*, 7(1), 7. <https://doi.org/10.1186/s40779-020-00238-8>
- Muslimah, M., Sulalah, I. A., Rohmawati, R., Sofi, N., & Zunairoh, Y. (2025). Best practice of summative assessment for Arabic writing proficiency: Evidence from Pondok Pesantren Modern Ar-Rahmah Kediri, Indonesia. *Asalibuna*, 8(02), 48–62. <https://doi.org/10.30762/asalibuna.v8i02.5221>
- Sagala, P. N., & Andriani, A. (2019). Development of higher-order thinking skills (HOTS) questions of probability theory subject based on Bloom's taxonomy. *Journal of Physics: Conference Series*, 1188, 012025. <https://doi.org/10.1088/1742-6596/1188/1/012025>

- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2018). How to design a useful test. In *Understanding medical education* (pp. 275–289). Wiley. <https://doi.org/10.1002/9781119373780.ch20>
- Seyaningsih, N., & Sari, W. A. K. (2021). Development of the assessment model based on order thinking skills (HOTS) to measure students' critical thinking. *Journal of Physics: Conference Series*, 1776(1), 012022. <https://doi.org/10.1088/1742-6596/1776/1/012022>
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Yang, F., & Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education*, 123, 97–108. <https://doi.org/10.1016/j.compedu.2018.04.006>
- Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, 14. <https://doi.org/10.3389/fpsyg.2023.1181712>