

Date Received : July 2026
Date Revised : August 2026
Date Accepted : August 2026
Date Published : August 2026

THE QUALITY OF ISLAMIC RELIGIOUS EDUCATION TEST ITEMS IN VOCATIONAL HIGH SCHOOLS

Mifta Ghina Dzakiya

UIN Sunan Gunung Djati Bandung, Indonesia (dzakiyaghina@gmail.com)

Tarsono

UIN Sunan Gunung Djati Bandung, Indonesia (tarsono@uinsgd.ac.id)

Zahra Uswah Hasanah

UIN Sunan Gunung Djati Bandung, Indonesia (zhr.uswah@gmail.com)

Ahmad Abdul Majid

UIN Sunan Gunung Djati Bandung, Indonesia (ahmadmandor300102@gmail.com)

Dadan Rusmana

UIN Sunan Gunung Djati Bandung, Indonesia (dadan.rusmana@uinsgd.ac.id)

Kata Kunci:	ABSTRAK
Asesmen; Butir Soal; Validitas	<p>Evaluasi hasil belajar melalui penilaian sumatif memegang peranan krusial dalam memotret pencapaian kompetensi siswa secara akurat. Namun, praktik di lapangan menunjukkan bahwa instrumen tes sering kali disusun tanpa melalui pengujian psikometris yang cermat, sehingga memicu terjadinya distorsi pengukuran. Penelitian ini bertujuan untuk memetakan distribusi butir soal yang valid dan tidak valid, serta mengidentifikasi akar penyebab ketidakvalidan instrumen ditinjau dari dimensi materi, konstruksi, dan keterbacaan bahasa. Metode penelitian yang digunakan adalah pendekatan kuantitatif deskriptif yang dipadukan dengan analisis isi kualitatif. Pengumpulan data dilakukan melalui dokumentasi naskah tes sumatif mata pelajaran Pendidikan Agama Islam dan Budi Pekerti (PAI-BP) untuk Kelas XI jurusan Teknik Kendaraan Ringan (TKR) di SMK Negeri Karangpucung, serta data jawaban dari 35 siswa. Data ditabulasi ke dalam matriks biner (skor 1 dan 0) dan dianalisis secara kuantitatif menggunakan rumus Excel dengan bantuan perangkat lunak SPSS, kemudian dibandingkan dengan nilai r-tabel sebesar 0,334. Hasil penelitian menunjukkan bahwa dari 30 butir soal pilihan ganda yang diuji, 14 butir dinyatakan valid dan 16 butir dinyatakan tidak valid atau harus digugurkan, termasuk adanya anomali berupa dua butir soal yang menghasilkan koefisien korelasi negatif (minus). Akar penyebab ketidakvalidan tersebut diidentifikasi meliputi pengecoh (distraktor) yang tidak homogen, pokok soal (stem) yang bias, kesalahan pengetikan, serta stimulus teks yang terlalu panjang sehingga memicu kelelahan membaca pada siswa SMK. Implikasi penelitian ini menekankan pentingnya bagi pendidik untuk melakukan rekonstruksi menyeluruh terhadap bank soal yang bermasalah serta menyusun pertanyaan yang lebih ringkas dan aplikatif demi menjamin akuntabilitas dan keadilan dalam sistem penilaian sekolah.</p>

Keywords:	ABSTRACTS
Assesment; Test Item; Validity	<p>Learning outcomes evaluation through summative assessment plays a crucial role in accurately capturing student competency achievement. However, field practices indicate that test instruments are frequently developed without meticulous psychometric testing, thereby triggering measurement distortion. This study aims to map the distribution of valid and invalid test items, as well as to identify the root causes of instrument invalidity across the dimensions of content, construction, and language readability. The research method employed is a descriptive quantitative approach integrated with qualitative content analysis. Data collection was conducted through documentation of the summative test manuscript for the Islamic Religious Education and Character (PAI-BP) subject of Class XI in the Light Vehicle Engineering (TKR) department at SMK Negeri Karangpucung, along with the response data from 35 students. The data were tabulated into a binary matrix (scores of 1 and 0) and analyzed quantitatively using Excel formulas assisted by SPSS software, then compared against an r-table value of 0.334. The results demonstrated that out of the 30 multiple-choice items tested, 14 items were declared valid and 16 items were declared invalid or discarded, including an anomaly of two test items yielding negative (minus) correlation coefficients. The root causes of invalidity were identified as non-homogeneous distractors, biased stems, typographical errors, and overly lengthy text stimuli that induced reading fatigue among vocational students. The implications of this study emphasize the urgency for educators to conduct a total reconstruction of flawed item banks and to formulate more concise, applicable questions to uphold accountability and fairness within the school assessment system.</p>

A. INTRODUCTION

Learning evaluation serves as an integral component within the educational system, functioning to measure the achievement of learning objectives in an objective and accountable manner. High-quality assessment instruments, particularly multiple-choice questions in summative assessments, must adhere to standardized psychometric criteria, notably empirical validity, to ensure that the resulting measurement outcomes can be scientifically justified. Within the framework of Classical Test Theory (CTT), item validity refers to the capacity of each individual item to accurately measure the specific construct it intends to measure, which is operationally demonstrated by an item-total correlation coefficient (point-biserial correlation) that exceeds the agreed critical threshold value (Artosandi 2025). In the subject of Islamic Religious Education and Character (PAI-BP), instrument validity becomes increasingly paramount given that exam questions are designed to measure not only factual knowledge but also the multidimensional understanding and internalization of religious values (Darmayanti, 2019).

However, reality in the field frequently reveals that the quality of PAI-BP assessment instruments at the Vocational High School (SMK) level is still far from meeting these ideal standards. Data from the Indonesian Central Bureau of Statistics (BPS) in 2024 recorded that the open unemployment rate for SMK graduates reached 8.62%, the highest compared to other educational levels, with one of the contributing factors being the low relevance and accuracy of the competency assessment system in vocational schools. (Zuhri et al., 2023). This condition indicates that the evaluation system in vocational high schools (SMK), including normative subjects such as Islamic Religious Education and Character (*PAI-BP*), is not yet fully capable of producing assessment data that are valid and representative of the students' actual abilities. (Yolnasdi 2025). The weak validation process of instruments prior to their utilization in summative examinations constitutes a root problem that has yet to receive adequate attention, particularly in vocational high schools (*SMK*) within the engineering department, where students possess specific cognitive and psycholinguistics characteristics. (Susilana & Rosliana, 2025).

A number of prior studies have examined the quality of test items within the contexts of Islamic education and vocational education. Specifically, a study published in *Al-Tanzim Journal* by (Jamalia, Huriyah, & Yuliati, 2022) found that the Islamic Religious Education (PAI) exam instruments in schools demonstrated significantly varied levels of validity depending on the teachers' pedagogical competence, particularly in their understanding of standardized test instrument construction techniques. Furthermore, a study in the *Jurnal Pendidikan Agama Islam* by (Chotija,dkk 2025) found that evaluation instruments in Islamic education must be capable of empirically integrating theological-religious values through valid psychometric testing. Furthermore, a study in the *Jurnal Pendidikan Agama Islam* by (Kadir & Umiarso, 2021)

found that the implementation of the learning system in Madrasah Aliyah revealed that limitations in teacher competence to formulate adaptive learning strategies had a direct impact on assessment bias, wherein the resulting learning evaluation outcomes were not yet capable of truly reflecting the actual capacity and ability of the students in the field. (Agustina & Yuana, 2025).

This study offers a more comprehensive and contextualized approach to evaluating the quality of Islamic Religious Education and Character (*PAI-BP*) test instruments. The novelty of this research lies in the integration of two analytical levels that have previously been conducted in isolation: first, a quantitative psychometric analysis utilizing point-biserial correlation to map the empirical validity of each test item; and second, a qualitative analysis through content analysis that critically examines the root causes of item invalidity across the dimensions of content alignment, distractor construction, and language readability, while considering the cognitive characteristics of vocational engineering students. This integrated approach positions this study as a novel methodological contribution to the field of Islamic Religious Education evaluation in Indonesia. (Anshari, dkk 2024).

This study offers a more comprehensive and contextualized approach to evaluating the quality of Islamic Religious Education and Character (*PAI-BP*) instruments. The novelty of this research lies in the integration of two analytical levels that have traditionally been conducted separately: first, a quantitative psychometric analysis utilizing point-biserial correlation Based on the aforementioned urgency, this study aims to: (1) describe the empirical validity quality of the Class XI *PAI-BP* multiple-choice items in the Light Vehicle Engineering (TKR) department at SMK Negeri Karangpucung based on point-biserial correlation analysis; and (2) identify the root causes of item invalidity through a qualitative analysis examining the alignment of content with learning objectives, the construction quality of distractors, and the level of language readability (Salamatov, dkk 2021). The results of this study are expected to provide theoretical contributions to the development of literature on psychometric-based Islamic Religious Education (*PAI*) evaluation, while simultaneously providing operational recommendations that can be implemented by educators to enhance the quality of assessment instruments within vocational education environments. (Rifka Alkhilyatul, dkk 2024).

B. METHOD

This study utilizes an evaluative descriptive quantitative approach integrated with qualitative content analysis of the test item drafts. This approach was selected due to its alignment with the research objectives, which aim to systematically describe the empirical quality of the PAI-BP assessment instrument and identify the root causes of item invalidity. The research locus is situated at SMK Negeri Karangpucung, Cilacap Regency, under the consideration that the test instruments serving as the object of study originated from and were officially utilized by the institution, thereby ensuring data authenticity and relevance. The research subjects comprise all active Class XI students majoring in Light Vehicle Engineering (TKR) during the even semester of the 2025/2026 academic year, totaling 35 respondents. These subjects were determined through a total sampling technique, considering the limited and homogeneous nature of the population within a single study group (Raihanan Syafiq, Alwi, & Zuhriah, 2024).

Data collection techniques were carried out through documentation studies of three sources, namely: (1) a draft of 30 multiple-choice questions along with the official answer key; (2) the Learning Objective Flow (Alur Tujuan Pembelajaran / ATP) document as a reference for content alignment; and (3) the response sheets of 35 students. Quantitative data analysis was conducted using the Point Biserial correlation technique r_{pbi} assisted by SPSS software, where response scores were binary-coded (correct = 1, incorrect = 0) and correlated with each student's total score. A test item is declared valid if the value of r_{pbi} is greater than the critical value to critically analyze the root causes of item invalidity based on the three standard pillars of test instrument construction, namely the alignment of content with learning objectives, the structural integrity of distractors, and the readability of the test item language (Raihanan Syafiq,dkk 2024).

The data analysis phase in this study is divided into two linear stages that mutually confirm each other. In the first stage, a quantitative parameter analysis is conducted, wherein the binary numerical data from the spreadsheet are processed using the point-biserial correlation formula r_{pbi} assisted by SPSS software. The applied correlation formula calculates the difference between the mean total score of students who answered correctly on a specific item and the mean total score of all test-takers, divided by the standard deviation of the total score, and then multiplied by the square root of the quotient of the proportion of students who answered correctly and the proportion of students who answered incorrectly on that item. This calculated r value is directly compared with the critical r table threshold value of 0.334, which is obtained through the degrees of freedom ($df = N - 2$ ($35 - 2 = 33$)) at a significance level (α) of 5% for a one-tailed test. Test items are categorized as valid if the $r_{hitung} > 0,334$ or yields a negative value (Sahin, 2025).

Subsequently, in the second stage, all test items declared invalid or discarded in the quantitative testing undergo a comprehensive and in-depth qualitative examination using content analysis techniques. This textual analysis refers to three dimensions of evaluation instrument quality standardization in a comprehensive manner. In the content aspect, the analysis is directed to examine the alignment degree between the substance of the questions and the competency indicators embedded within the *ATP* document (Sridharan 2025). In the construction aspect, the analysis focuses on diagnosing the mechanical function of the incorrect answer options (distractors), the homogeneity level of alternative choices, and the presence of double negative sentence structures that potentially induce cognitive bias in students. Lastly, in the language aspect, the evaluation is conducted on editorial readability, typographical errors (typos), the length of narrative text stimuli, and the alignment of psycholinguistic barriers with the specific psychological characteristics of vocational engineering students. This entire sequence of stages operates continuously to generate a holistic and precise conclusion (Rahdiyanta, dkk 2019).

C. RESULT AND DISCUSSION

Findings

Empirical validity testing was conducted on 30 multiple-choice test items of Islamic Religious Education and Character (PAI-BP) based on the response data of 35 Class XI students majoring in Light Vehicle Engineering (TKR) at SMK Negeri Karangpucung, utilizing the Point-Biserial correlation (r_{pbi}) technique assisted by SPSS software. The response scores were coded binarily (correct = 1, incorrect = 0) and correlated with the total score of each student. The decision-making criteria were established based on the $r = 0,334$ ($df = 33$, $\alpha = 5\%$, one tailed test), whereby a test item was declared valid if $r_{hitung} \geq 0,334$, and declared invalid if $r_{hitung} < 0,334$ or yielded a negative value.

The detailed results of the validity test for all test items are presented in Table 1 below.

Item Number	r count Value	r tabel Value	Item Quality Classification Criteria	Decision Status
Item 1	0.562	0.334	Moderate Validity	VALID
Item 2	0.115	0.334	Very Low Validity	INVALID
Item 3	0.485	0.334	Moderate Validity	VALID
Item 4	0.421	0.334	Moderate Validity	VALID
Item 5	0.612	0.334	High Validity	VALID

Item 6	0.554	0.334	Moderate Validity	VALID
Item 7	0.188	0.334	Very Low Validity	INVALID
Item 8	0.674	0.334	High Validity	VALID
Item 9	0.498	0.334	Validitas Cukup	VALID
Item 10	0.712	0.334	High Validity	VALID
Item 11	0.505	0.334	Moderate Validity	VALID
Item 12	0.214	0.334	Moderate Validity	INVALID
Item 13	0.442	0.334	Moderate Validity	VALID
Item 14	0.531	0.334	Moderate Validity	VALID
Item 15	-0.035	0.334	Negative Correlation (Severely Flawed)	INVALID
Item 16	0.41	0.334	Moderate Validity	VALID
Item 17	0.266	0.334	Low Validity	INVALID
Item 18	0.455	0.334	Moderate Validity	VALID
Item 19	0.143	0.334	Very Low Validity	INVALID
Item 20	0.385	0.334	Low Validity (Minimum Validity)	VALID
Item 21	0.29	0.334	Low Validity	INVALID
Item 22	0.102	0.334	Very Low Validity	INVALID
Item 23	0.311	0.334	Low Validity	INVALID
Item 24	0.085	0.334	Very Low Validity	INVALID
Item 25	0.225	0.334	Low Validity	INVALID
Item 26	0.195	0.334	Very Low Validity	INVALID
Item 27	-0.012	0.334	Negative Correlation (Severely Flawed)	INVALID
Item 28	0.24	0.334	Low Validity	INVALID
Item 29	0.165	0.334	Very Low Validity	INVALID
Item 30	0.05	0.334	Very Low Validity	INVALID

The decision-making criteria were established as follows:

1. If $r \text{ count} > 0,334$, the test item is declared Valid.
2. If $r \text{ count} < 0,334$ or yields a negative value ($-$), the test item is declared Invalid (Discarded).

Based on the decision-making criteria presented in Table 1, the empirical validity test results demonstrated a highly varied distribution of validity indices. Through a direct comparison between the $r \text{ count}$ output from SPSS and the critical threshold of r table sebesar $0,334$, the final results indicated that 14 test items (46,7%) were declared valid, while 16 test items (53.3%) were declared invalid or discarded. Specifically, the 14 items that successfully met the statistical threshold because their $r \text{ count} > 0,334$ were items number 1, 3, 4, 5, 6, 8, 9, 10, 11, 13, 14, 16, 18, and 20. Observed from their quality classification, the distribution of these valid items ranged from low to high validity. Item 10 recorded the highest index with a correlation coefficient of $r \text{ hitung} = 0,712$, proving that this item possesses a highly robust discrimination power in separating the cognitive capacities between the high-ability and low-ability student groups. Meanwhile, item 20 sat exactly at the minimum safe threshold of statistical passing with a coefficient of $r \text{ count} = 0,385$. Based on these parameters, the 14 valid test items are declared to be functioning mechanically at an optimal level and are highly eligible to be retained in the school's item bank (Saepuzaman, dkk 2025).

Conversely, the 16 test items declared invalid or discarded due to their coefficients falling below the threshold line of $0,334$ were items number 2, 7, 12, 15, 17, 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, and 30. The majority of these discarded items were trapped in the low and very low validity categories with correlation coefficients approaching zero, as extremely demonstrated by item 24 ($r \text{ count} = 0,085$) dan soal nomor 30 ($r \text{ count} = 0,050$). These remarkably low index values signify a loss of the instrument's measurement function due to random response patterns, representing a high intensity of speculative guessing by students during the examination.

The most significant correlation discrepancy was identified in item 15 ($r \text{ count} = -0,035$) and item 27 ($r \text{ count} = -0,012$) which yielded negative (minus) correlation coefficients. Theoretically, this inverse relationship proves the existence of a fatal dysfunction within the internal structure of the instrument, where the majority of the high-ability student group (*top total scores*) were misled into selecting the wrong options, whereas the low-ability student group (*bottom total scores*) unexpectedly benefited from random chance or biased stimulus phrasing. This empirical phenomenon of degraded negative correlation parameters and low validity coefficients across the 16 items serves as the scientific basis for a diagnosis through an in-depth qualitative content analysis in the Discussion section. (Yulastri et al., 2025)

Analysis/Discussion

1. Quantitative Parameter Analysis of PAI-BP Item Empirical Validity

The quantitative analysis in this study provides empirical confirmation that the summative exam instrument for Class XI PAI-BP at SMK Negeri Karangpucung is not yet in an ideal condition as a measurement tool for student learning outcomes. Based on the accumulated binary data processed, the fact that 16 test items (53.3%) suffered a mass failure represents a serious methodological gap in the planning and drafting phases of the questions by educators. Referring to the foundational concepts of psychometrics, instrument validity is essentially the degree of accuracy, correctness, and legitimacy of a measurement tool in executing its specific measuring function. When an evaluation instrument exhibits a failure rate exceeding half of the total material tested, the accumulated total scores obtained by students no longer purely reflect their true religious cognitive capacity. Instead, these scores have been massively distorted by measurement error triggered by technical factors outside the target variable intended to be measured.

If examined more closely based on the coefficient distribution in Table 1, the test items located within the invalid zone are predominantly trapped in low or very low correlation classifications that approach zero. This phenomenon is clearly observable in the statistical data of item 24 ($r_{\text{hitung}} = 0,085$) dan soal nomor 30 ($r_{\text{hitung}} = 0,050$). This highly weak relationship that deviates from linearity mechanically indicates that these test items have lost their formal measurement function as well as their discrimination parameters. In educational evaluation theory, a test item can only be deemed valid and contribute positively to test reliability if its correlation coefficient value surpasses the critical baseline of $r_{\text{tabel}} = 0,334$ pada taraf signifikansi 5% (Puspita & Sari, 2024).

When the coefficient falls below this threshold, the instrument loses its sensitivity to filter and distinguish between truly competent students (the upper group) and students who have not mastered the material (the lower group). The random distribution of response data on these items acts as a strong indicator of mass speculative guessing behaviors among the students. Because the item stimuli fail to stimulate the students' logical reasoning, students ultimately resort to speculative decisions when choosing answer options, which automatically invalidates the scientific integrity of the evaluation tool.

The most crucial correlation discrepancy requiring careful attention was found in item 15 ($r_{\text{hitung}} = -0,035$) dan soal nomor 27 ($r_{\text{hitung}} = -0,012$) which generated negative coefficients. Theoretically, a negative Point-Biserial correlation value indicates an extreme directional reversal dysfunction within the item's internal structure (Krisnawati, Artanti, & Umar, 2024). This minus figure serves as authentic evidence in the field that the test items operate inversely to a fair assessment function. Ironically, a negatively correlated item penalizes the upper-

group students who possess analytical critical reasoning because the majority are misled into selecting options designated as incorrect by the official answer key. Conversely, the lower-group students who answer speculatively without the burden of complex reasoning benefit from random coincidence or biased option phrasing. The emergence of this negative data acts as a scientific alarm that the quality of the sentences drafted by the educator suffers from severe structural flaws, making an *a priori* qualitative validation through a complete reconstruction of the draft instruments an absolute academic urgency that cannot be deferred.

2. Qualitative Analysis of Causal Factors Underlying Test Item Invalidity

In the group of items categorized under "Low Validity" with coefficients close to the threshold line (items 12, 17, 21, 23, 25, 28), editorial tracking indicates that the primary problem lies within the structural construction, specifically the dysfunction of non-homogeneous distractors. In these items, the remaining positive direction of the correlation shows that the questions are fundamentally still capable of guiding high-ability students to answer correctly. However, their discrimination power is weakened because one or two distractor options are factually too easy to eliminate. Consequently, some medium-to-low ability students were able to select the correct answer through a simple elimination process rather than genuine concept mastery. Explained that a response distribution influenced by non-homogeneous distractors serves as a strong indicator of partial speculative guessing a pattern entirely consistent with the findings in this specific cluster of items (Magdalena 2021).

In the cluster of items categorized as "Very Low Validity" with coefficients approaching zero (items 2, 7, 19, 22, 24, 26, 29, 30), text penelusuran reveals a more fundamental problem regarding content, namely the misalignment between the item indicators and the actual learning outcomes taught within the ATP. A coefficient close to zero theoretically implies that the probability of answering correctly is relatively equal between high-ability and low-ability students. This pattern is highly consistent with a condition where the material tested falls completely outside the scope mastered by both groups of students. As a result, the answers provided tend to be purely random for both ability levels, rather than merely being influenced by weak distractors.

From Item 15 ($r = -0,035$) and item 27 ($r = -0,012$) which produced negative coefficients, editorial tracking indicates flaws within the question stems that potentially generate double interpretations. Consequently, students who read carefully and analytically were directed to select options designated as incorrect by the official answer key. Explained that a negative point-biserial correlation coefficient indicates an inverse relationship dysfunction within the internal structure of the question. Under this condition, upper-group students who consistently answer correctly on the majority of other items are misled by a specific item, while lower group students maintain an equal probability of coincidentally

picking the option designated as the key. This finding confirms that the root cause of both items lies in the structural quality of the phrasing rather than the students' mastery of the material. This distinction is critical because it alters the remediation strategy: items 15 and 27 require an extensive linguistic and redaksional revision rather than a replacement of the learning content (Krisnawati et al., 2024).

Regarding the linguistic aspect, the study also identified the utilization of relatively lengthy narrative text stimuli across several discarded items, particularly those testing contextual understanding. For TKR department students whose daily academic routines interact heavily with brief and procedural work instructions, exposure to dense paragraph stimuli potentially induces reading fatigue, which subsequently reduces meticulousness in identifying the core question, as suggested by Agustina and Yuana (2025). However, given that not all 16 discarded items featured lengthy text stimuli, this element is more accurately viewed as a contributing factor that exacerbated the impact of structural flaws in certain items, rather than a single cause applicable uniformly to all discarded items.

Consequently, this tracking yields a highly precise causal mapping: the low validity cluster (bordering the threshold) is dominated by distractor construction issues; the very low validity cluster (approaching zero) is dominated by content misalignment with the ATP; whereas the two negatively correlated items are dominated by redaksional issues within the question stems (Putri, Kuhon, & Palandeng, 2024).

3. Theoretical and Practical Implications

Theoretically, these findings reinforce and refine the arguments regarding the importance of item analysis before instruments are deployed, demonstrating that instrument validity failure is not a monolithic phenomenon but consists of several structural sub-types with distinct root causes (Ninghardjanti 2021). This study also enriches the literature on PAI evaluation by demonstrating that the psychological characteristics of students in specific vocational majors can act as a complicating contextual factor, though not the primary cause, of validity failure in certain items, aligning with the findings of Rifka Alkhilyatul et al. (2024) regarding the importance of student context in test instrument validation. Practically, this mapping enables PAI-BP teachers to formulate layered remediation strategies:

1. Items within the near-threshold zone (17, 21, 23, and with lower priority 12, 25, 28) can be sufficiently rehabilitated by replacing the non-homogeneous distractor options.
2. Items with coefficients close to zero (2, 7, 19, 22, 24, 26, 29, 30) require a rigorous review of their content alignment with the ATP and most likely need to be replaced with entirely new items.
3. Items with negative correlations (15 and 27) demand a comprehensive linguistic restructuring of the question stems, accompanied by a readability check before

being redeployed.

Furthermore, these results underscore the urgent need to establish collaborative item review forums (either through MGMP or internal school networks) and to implement limited trial testing before items are utilized as official summative instruments, ensuring that such structural flaws can be intercepted during the developmental stage.

CONCLUSION

Based on the results of the analysis and discussion, it can be concluded that the quality of the summative exam instrument for the Class XI PAI-BP subject in the TKR department at SMK Negeri Karangpucung does not yet fully meet ideal psychometric standards. Out of the 30 multiple-choice items evaluated, only 14 items (46.7%) were declared valid, while the remaining 16 items (53.3%) failed due to r count values falling below the r_{table} threshold of 0.334. This high percentage of invalid items was qualitatively identified to stem from structural weaknesses, such as non-homogeneous distractors, the appearance of negative correlation anomalies in items 15 and 27 due to biased stems, and excessive narrative text burdens that induced reading fatigue among vocational students. Therefore, this study provides a strong recommendation for educators to secure the 14 valid questions into the school's item bank and to conduct a total reconstruction of the 16 discarded questions by designing more logical distractors and formulating question stems that are more concise, dense, and applicable to uphold accountability and fairness within the school's learning assessment system.

REFERENCES

- Agustina, I. P., & Yuana, O. A. (2025). Peran Evaluasi Dalam Meningkatkan Kualitas Pendidikan Agama Islam Di Era Digital. *Al-Fatih: Jurnal Pendidikan Dan Keislaman*, 8(1), 250–272.
- Anshari, M. I., Nasution, R., Irsyad, M., Alifa, A. Z., & Zuhriyah, I. A. (2024). Analisis Validitas Dan Reliabilitas Butir Soal Sumatif Akhir Semester Ganjil Mata Pelajaran Pai. *Edukatif: Jurnal Ilmu Pendidikan*, 6(1), 964–975. <https://doi.org/10.31004/Edukatif.V6i1.5931>
- Artosandi, Y. S. R., Prijowuntato, S. W., & Abdul Kadir, N. B. (2025). Empowering Career Exploration: Designing A Vocational Interest Instrument For Junior High School Students. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 29(2), 154–166. <https://doi.org/10.21831/Pep.V29i2.84561>
- Başer, A., Sönmez, Ö. F., Kürklü Arpaçay, D., & Şahin, H. (2025). Adapting Organizational Culture Scale Into Healthcare Professional Education: A Scale Validity And Reliability Analysis. *Human Resources For Health*, 23(1). <https://doi.org/10.1186/S12960-025-01006-2>
- Chotijah, H., Suparman, & Kusumaningtyas, D. A. (2025). Beyond Technocratic Competence: A Psychometric Validation Of A Values-Embedded Teacher Assessment Model In Indonesian Islamic Schools. *Jurnal Pendidikan Agama Islam*, 22(2), 585–604. <https://doi.org/10.14421/Jpai.V22i2.12965>
- Darmayanti, H. (2019). Evaluasi Pendidikan. *Jurnal Edukatif*, V(1), 58–65.
- Jamalia, J., Huriyah, L., & Yuliati, W. (2022). E-Report Application: The Role Of Stakeholders In The Assessment System In Schools. *Al-Tanzim: Jurnal Manajemen Pendidikan Islam*, 6(4), 999–1009. <https://doi.org/10.33650/Al-Tanzim.V6i4.3112>
- Kadir, A., & Umiarso. (2021). Evaluation Of The Application Of Distance Learning Programs In Madrasah Aliyah Malang Regency. *Jurnal Pendidikan Agama Islam*, 18(1), 109–124. <https://doi.org/10.14421/Jpai.2021.181-06>
- Krisnawati, E., Artanti, K. D., & Umar, N. H. (2024). Uji Validitas Dan Reliabilitas Instrumen Penelitian Dukungan Suami Terhadap Hambatan Penggunaan Metode Kontrasepsi Jangka Panjang Pada Multipara Akseptor Aktif Di Surabaya. *Media Gizi Kesmas*, 13(2), 659–664. <https://doi.org/10.20473/Mgk.V13i2.2024.659-664>
- Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal. *Bintang: Jurnal Pendidikan Dan Sains*, 3(2), 198–214. Retrieved From <https://ejournal.stitpn.ac.id/index.php/Bintang>
- Ninghardjanti, P., & Dirgatama, C. H. A. (2021). Building Critical Thinking Skills Through A New Design Mobile-Based Interactive Learning Media Knowledge Framework. *International Journal Of Interactive Mobile Technologies*, 15(17), 49–68. <https://doi.org/10.3991/Ijim.V15i17.23801>

- Puspita, A., & Sari, I. (2024). Validitas Dan Reliabilitas Soal Ujian Sumatif Tengah Semester (Sts) Kelas X Fase-E Sma Negeri 8 Palembang Tahun Pelajaran 2023 / 2024. *Alfabeta: Jurnal Bahasa, Sastra, Dan Pembelajarannya*, 7(2), 337-347.
- Putri, M. A., Kuhon, F. V., & Palandeng, H. M. F. (2024). Uji Validitas Dan Reliabilitas Instrumen Penelitian. *Jurnal Kedokteran Komunitas Dan Tropik*, 12(2), 635-640. <https://doi.org/10.35790/jkkt.v12i2.59634>
- Rahdiyanta, D., Nurhadiyanto, D., & Munadi, S. (2019). The Effects Of Situational Factors In The Implementation Of Work-Based Learning Model On Vocational Education In Indonesia. *International Journal Of Instruction*, 12(3), 307-324. <https://doi.org/10.29333/iji.2019.12319a>
- Raihanan Syafiq, S. N., Alwi, J. L., & Zuhriah, I. A. (2024). Teknik Analisis Validitas Tes Menggunakan Spss Dalam Mengukur Kemampuan Hasil Belajar Materi Ski Kelas Xi Sma. *Afeksi: Jurnal Penelitian Dan Evaluasi Pendidikan*, 5(6), 1020-1030. <https://doi.org/10.59698/afeksi.v5i6.388>
- Rifka Alkhilyatul Ma'rifat, I Made Suraharta, I. I. J. (2024). Analisis Validasi Soal Tes Hasil Belajar Pada Pelaksanaan Pembelajaran Bahasa Arab Di Pusat Pengembangan Bahasa (P3b) Uin Suska Riau. 2, 306-312.
- Saepuzaman, D., Istiyono, E., Haryanto, & Retnawati, H. (2025). Characteristics Of Two-Tier Multiple Choice (Ttmc) High Order Thinking Skill (Hots) Instruments And The Estimation Of Hots Abilities Of Vocational Education Students Using The Item Response Theory Approach. *Journal Of Engineering Education Transformations*, 38(Special Issue), 13-22. <https://doi.org/10.16920/jeet/2025/V38is3/25086>
- Salamatov, A., Gafarova, E., Belevitin, V., Gafarov, M., & Gordeeva, D. (2021). Machine Learning Methods And Qualimetric Approach To Determine The Conditions For Train Students In The Field Of Environmental And Economic Activities. *International Journal Of Emerging Technologies In Learning*, 16(3), 72-85. <https://doi.org/10.3991/ijet.v16i03.17715>
- Sridharan, K., & Sivaramakrishnan, G. (2025). Angoff Methods In Standard Setting In Health Professional Education: A Systematic Review And Meta-Analysis. *Bmc Medical Education*, 25(1). <https://doi.org/10.1186/s12909-025-08300-6>
- Susilana, R., & Rosliana, L. (2025). Curriculum Evaluation In Tourism Education : Assessing The Fulfillment Of Obe Principles. 29(2), 240-255.
- Yolnasdi, Simatupang, W., Sukardi, & Fadillah, R. (2025). Effectiveness Of Project-Based Cooperative Learning Model In Electrical Installation Practical Courses. *Salud, Ciencia Y Tecnologia*, 5. <https://doi.org/10.56294/saludcyt20251779>
- Yulastri, A., Ganefri, Fiandra, Y. A., Ferdian, F., Elfizon, Samala, A. D., ... Zahari, M. S. M. (2025). Building The Next Generation Of Digital Entrepreneurs: Lessons From Indonesian Vocational Students. *Salud, Ciencia Y Tecnologia*, 5. <https://doi.org/10.56294/saludcyt20252111>

Zuhri, A., Ramírez-Coronel, A. A., Al-Hawary, S. I. S., Dwijendra, N. K. A., Muda, I., Pallathadka, H., ... Sunarsi, D. (2023). Evaluation Of The Role Of Islamic Lifestyle In Communication Skills Of Muslim Couples. *Hts Teologiese Studies / Theological Studies*, 79(1), 1–6. <https://doi.org/10.4102/Hts.V79i1.8185>

