E-ISSN: 2962-6110
P-ISSN: 2964-7843
JSRET
Journal of Scientific Research, Education, and Technology
jsret.knpub.com

Quality of Scholastic Tests in the New Student Admission Selection *(SPMB)* at Universitas Negeri Surabaya in the 2023 Academic Year

Tri Rijanto, Edy Sulistyo, Joko, Puput Wanarti Rusimamto
Faculty of Engineering, State University of Surabaya, East Java, Indonesia

ABSTRACT

Scholastic tests play a critical role in university admission systems, serving as gatekeeping mechanisms that influence both institutional quality and educational equity. Despite their widespread use, empirical evaluations of institutional admission test quality remain limited, particularly in developing country contexts. This study aimed to comprehensively evaluate the psychometric quality of the Scholastic Potential and Basic Ability Test (SPMB) administered at Universitas Negeri Surabaya during the 2023 academic year, examining reliability, item difficulty, discrimination indices, and distractor effectiveness. A quantitative descriptive research design using ex-post facto analysis was employed. The study analyzed test response data from 270 candidates who completed the 45-item SPMB, consisting of three subtests: Verbal Ability (15 items), Numerical and Reasoning Ability (15 items), and Figural Comprehension Ability (15 items). Data analysis utilized Classical Test Theory frameworks, calculating Kuder-Richardson Formula 20 (KR-20) reliability coefficients, item difficulty indices (p-values), point-biserial discrimination coefficients (rpbis), upper-lower 27% discrimination indices (D), and distractor effectiveness metrics using SPSS 26.0 and ITEMAN 4.3 software. The total test demonstrated good internal consistency reliability (KR-20 = 0.84) with a mean score of 25.84 (SD = 6.78, 57.42% of maximum). Approximately 62% of items exhibited optimal moderate difficulty ($0.40 \leq p < 0.80$), and 73% demonstrated good-to-excellent discrimination (rpbis ≥ 0.30). However, three items showed poor discrimination (rpbis < 0.20), 22 distractors were non-functional (16.30%), and six distractors exhibited problematic positive discrimination (4.44%). Subtest reliabilities ranged from 0.70 to 0.75, classified as acceptable. The SPMB demonstrated generally satisfactory psychometric quality but requires targeted improvements through systematic item revision, enhanced item writer training, and continuous quality monitoring. Findings provide actionable guidance for evidence-based test refinement and contribute empirical evidence to admission testing literature in Southeast Asian higher education contexts.

Keywords: *Admission Testing, Psychometric, Test Reliability, Item Analysis, Classical Test Theory*

Corresponding author
Name: Tri Rijanto
Email: Tririjanto@unesa.ac.id

INTRODUCTION

The quality of scholastic tests in higher education admission systems has long been recognized as a critical determinant of institutional excellence and educational equity

(Mountford-Zimdars, 2018). Admission tests serve as gatekeeping mechanisms that not only predict academic success but also shape the demographic and intellectual composition of universities (Kuncel & Hezlett, 2010). In many countries, standardized scholastic tests have become indispensable tools for selecting candidates who demonstrate the cognitive abilities, subject mastery, and academic potential necessary for tertiary education (Camara & Echternacht, 2000). However, the validity and reliability of these instruments remain subjects of ongoing scholarly debate, particularly concerning their capacity to ensure fairness across diverse populations and their alignment with institutional missions (Sackett et al., 2009).

From a psychometric perspective, test quality is fundamentally determined by several interrelated properties, including validity, reliability, item difficulty, and discrimination indices (Downing, 2004). Validity refers to the extent to which a test measures what it purports to measure and whether inferences drawn from test scores are appropriate for their intended purpose (Jeffrey, 2017). Reliability, conversely, concerns the consistency and stability of test scores across different occasions, forms, or raters (Feldt & Brennan, 1989). Beyond these classical indicators, item-level characteristics such as difficulty (the proportion of examinees answering correctly) and discrimination (the ability of an item to differentiate between high and low achievers) provide essential information about individual test components (Brookhart & McMillan, 2019). Item Response Theory (IRT) has further advanced our understanding of test quality by modeling the relationship between examinee ability and item characteristics, offering more sophisticated approaches to test construction and evaluation (Hambleton & Swaminathan, 2013). Contemporary assessment scholarship emphasizes that high-quality admission tests should not only demonstrate strong psychometric properties but also exhibit fairness, transparency, and alignment with institutional learning outcomes (Vahrenhold & Paul, 2014).

In developing countries, university admission systems often face unique challenges related to resource constraints, large applicant pools, and diverse educational backgrounds among candidates (Glewwe & Kremer, 2006). These contextual factors can complicate the design and implementation of high-quality scholastic tests. While standardized national examinations exist in many nations, individual institutions frequently develop their own selection instruments to assess program-specific competencies or to complement centralized testing systems (Kellaghan & Greaney, 2019). However, institutional admission tests in developing contexts are not always subjected to rigorous psychometric evaluation, potentially compromising their effectiveness as selection tools. Research has documented instances where admission tests lack adequate validation evidence, exhibit problematic item characteristics, or fail to predict academic performance reliably (Burton & Ramist, 2001). Such deficiencies can undermine the meritocratic principles underlying university admissions and may inadvertently disadvantage certain groups of applicants (Alon & Tienda, 2007).

Universitas Negeri Surabaya (UNESA), one of Indonesia's leading state universities specializing in education and teacher training, conducts its own Scholastic Potential and Basic Ability Test (Seleksi Penerimaan Mahasiswa Baru or SPMB) to select prospective

students beyond the national selection pathways. The SPMB at UNESA is designed to assess candidates' scholastic abilities across multiple domains, including verbal reasoning, quantitative reasoning, and subject-specific knowledge relevant to various study programs. In the 2023 academic year, the SPMB served as a critical component of the university's admission strategy, selecting students who would contribute to UNESA's mission of producing qualified educators and professionals. Given the high-stakes nature of this examination and its implications for both institutional quality and individual educational trajectories, systematic evaluation of the SPMB's psychometric properties becomes imperative.

Despite the widespread use of institutional scholastic tests in university admissions, empirical studies examining their quality remain surprisingly limited, particularly in Southeast Asian contexts (Zainuddin et al., 2020). While considerable research has focused on national standardized examinations, institution-specific admission tests have received comparatively less scholarly attention (Nguyen et al., 2018). This gap is particularly pronounced regarding comprehensive psychometric analyses that integrate classical test theory and item-level evaluation to provide actionable insights for test improvement (Ackerman et al., 2022). Furthermore, few studies have systematically documented the quality of admission tests at teacher education universities, where the stakes are arguably higher given the multiplier effect of teacher quality on educational systems (Claudia-Nicoleta Păun & Adrian Costea, 2025).

This study addresses these gaps by conducting a comprehensive evaluation of the SPMB scholastic test quality at Universitas Negeri Surabaya for the 2023 academic year. The primary objective is to assess the psychometric properties of the SPMB, including test reliability, item difficulty distribution, item discrimination indices, and the overall validity of the instrument as an admission selection tool. By providing empirical evidence regarding test quality, this study aims to inform evidence-based refinements to UNESA's admission testing practices and contribute to the broader literature on institutional admission assessment in developing country contexts. The findings are expected to offer practical insights for test developers and institutional policymakers seeking to enhance the fairness, accuracy, and effectiveness of their selection instruments.

METHOD

Research Design

This study employed a quantitative descriptive research design using ex-post facto analysis to evaluate the psychometric quality of the Scholastic Potential and Basic Ability Test (SPMB) administered at Universitas Negeri Surabaya during the 2023 academic year. The ex-post facto approach was deemed appropriate as it allowed for the examination of test quality characteristics after the administration of the examination, without manipulation of variables (Kerlinger, 1986). The research utilized archival test data and examinee responses to conduct comprehensive psychometric analyses based on Classical Test Theory (CTT) principles (Allen & Yen, 1979). This design facilitated the assessment of test reliability, item difficulty, item discrimination, and distractor effectiveness, providing

empirical evidence regarding the technical quality of the admission instrument (Crocker & Algina, 2008).

Subjects/Population and Sample

The population of this study consisted of all candidates who participated in the SPMB at Universitas Negeri Surabaya during the 2023 academic year admission cycle. The total population comprised 312 test-takers who completed the scholastic test across multiple testing sessions conducted between May and June 2023. From this population, a sample of 270 examinees was selected for analysis after data cleaning procedures to ensure data quality and completeness. The sampling approach employed convenience sampling based on data availability and completeness criteria (Fraenkel, J. R., Wallen, N. E., & Hyun, 2017).

The SPMB examination consisted of 45 multiple-choice items distributed across three subtests: (1) Verbal Ability (15 items), assessing vocabulary, reading comprehension, and verbal reasoning skills; (2) Numerical and Reasoning Ability (15 items), measuring numerical competence, mathematical reasoning, and logical problem-solving skills; and (3) Figural Comprehension Ability (15 items), evaluating spatial reasoning, pattern recognition, and visual-spatial intelligence. All examinees completed the identical test form under standardized conditions in computer-based testing facilities at UNESA's main campus.

Data quality checks were performed to identify and address missing responses, duplicate records, or technical anomalies in the dataset. A total of 42 incomplete test records (13.46% of the population) were excluded from analysis due to premature test termination, excessive missing responses (>10% of items unanswered), or technical failures during test administration, resulting in a final analytical sample of 270 valid cases (86.54% of the total population).

Data Collection Procedure

Data collection was conducted through the university's official computer-based testing (CBT) system, which automatically recorded examinee responses, response times, and demographic information. The SPMB was administered over a one-week period in May 2023 across 6 testing sessions to accommodate the applicants. Each testing session lasted 90 minutes, with examinees receiving standardized written and verbal instructions regarding test procedures, scoring methods, and academic integrity policies.

Prior to data extraction, formal permission was obtained from UNESA's Admission Committee and the university's Research Ethics Board to ensure compliance with institutional research protocols and data privacy regulations. All personally identifiable information was removed from the dataset following ethical guidelines for secondary data analysis (H Kara, 2014). Each examinee was assigned a unique anonymous identifier to maintain confidentiality while enabling tracking of responses.

The raw data extraction process involved retrieving the following information from the CBT database: (1) item-level response matrices (correct/incorrect/omitted) for all 45 items across all examinees; (2) total test scores and subtest scores for each participant; (3)

item keys and scoring rubrics; (4) distractor selection frequencies for each multiple-choice option; and (5) relevant demographic variables. The complete dataset was exported to Microsoft Excel format and subsequently imported into statistical software for analysis.

Data Analysis

Data analysis was conducted using a multi-stage approach integrating descriptive statistics and classical psychometric indices. All statistical analyses were performed using SPSS Statistics version 26.0 (IBM Corp., 2019) and ITEMAN 4.3 software (Assessment Systems Corporation, 2014), specialized programs for item and test analysis based on Classical Test Theory frameworks.

Test reliability was evaluated using Kuder-Richardson Formula 20 (KR-20), the appropriate internal consistency coefficient for dichotomously scored items (Kuder & Richardson, 1937). The KR-20 coefficient was calculated for the total test and for each subtest independently. Reliability coefficients were interpreted using Nunnally and Bernstein's (1994) criteria: values ≥ 0.70 indicating acceptable reliability, ≥ 0.80 indicating good reliability, and ≥ 0.90 indicating excellent reliability. The Standard Error of Measurement (SEM) was also computed to estimate the precision of individual test scores using the formula: SEM = SD√(1-reliability), where SD represents the standard deviation of test scores.

Item difficulty (p-value) was calculated as the proportion of examinees answering each item correctly, ranging from 0.00 (no one answered correctly) to 1.00 (everyone answered correctly). Following conventional CTT standards, items were classified as: very difficult ($p < 0.20$), difficult ($0.20 \leq p < 0.40$), moderate ($0.40 \leq p < 0.80$), easy ($0.80 \leq p < 0.90$), and very easy ($p \geq 0.90$) (Garvin & Ebel, 1980). The optimal distribution of item difficulties was evaluated against the recommended guideline that the majority of items should fall within the moderate range (0.30-0.70) to maximize test information and discrimination (John, 2015).

Item discrimination was assessed using the point-biserial correlation coefficient (rpbis), which measures the relationship between performance on individual items and total test scores (Henrysson, 1971). The point-biserial correlation was calculated using SPSS, with values interpreted according to Ebel and Frisbie's (1991) classification: excellent discrimination ($rpbis \geq 0.40$), good discrimination ($0.30 \leq rpbis < 0.40$), acceptable discrimination ($0.20 \leq rpbis < 0.30$), and poor discrimination ($rpbis < 0.20$). Items with negative discrimination indices were flagged for review as they indicate problematic functioning where low-ability examinees outperform high-ability examinees (Haladyna et al., 2002).

Additionally, the discrimination index (D) was computed using the upper-lower 27% method, comparing the performance of high-scoring examinees (upper 27%) with low-scoring examinees (lower 27%) on each item (Kelley, 1939). The discrimination index was calculated as: $D = (U - L) / n$, where U is the number of correct responses in the upper group, L is the number of correct responses in the lower group, and n is the number of examinees

in each group. Discrimination indices were interpreted as: excellent ($D \geq 0.40$), good ($0.30 \leq D < 0.40$), fair ($0.20 \leq D < 0.30$), and poor ($D < 0.20$) (Ebel & Frisbie, 1991).

The effectiveness of distractors (incorrect answer options) was evaluated by examining the selection frequency of each option and calculating the point-biserial correlation between selecting each distractor and total test scores. Effective distractors should be selected by some examinees (typically at least 5%) and should exhibit negative point-biserial correlations, indicating that lower-ability examinees are more likely to select them than higher-ability examinees (Haladyna & Downing, 1993). Distractors selected by fewer than 5% of examinees were identified as non-functional and recommended for revision or replacement.

Comprehensive descriptive statistics were calculated for the total test and each subtest, including mean scores, standard deviations, minimum and maximum scores, skewness, and kurtosis. Score distributions were examined using histograms and normality tests (Kolmogorov-Smirnov and Shapiro-Wilk tests) to assess whether score distributions approximated normal distributions, which is desirable for norm-referenced admission tests (Crocker & Algina, 1986). Ceiling and floor effects were evaluated by examining the percentage of examinees scoring at the highest and lowest possible score ranges.

All statistical tests were conducted using a significance level of $\alpha = 0.05$. Descriptive findings and psychometric indices were tabulated and presented with appropriate visual representations including frequency distributions, histograms, and scatter plots where applicable.

FINDING AND DISCUSSION

RESEARCH RESULT

The results of the psychometric analysis of the SPMB scholastic test at Universitas Negeri Surabaya for the 2023 academic year are presented in four main sections: (1) descriptive statistics and score distributions, (2) test reliability, (3) item difficulty analysis, and (4) item discrimination and distractor effectiveness. These findings provide comprehensive empirical evidence regarding the technical quality of the admission instrument.

Descriptive Statistics and Score Distribution

Table 1 presents the descriptive statistics for the total SPMB test and its three subtests based on the analysis of 270 valid test records. The mean total score was 25.84 out of 45 possible points (57.42%), with a standard deviation of 6.78, indicating moderate variability in examinee performance. The minimum score observed was 11, while the maximum score was 40, suggesting that no examinee answered all items correctly or performed at chance level across the entire test.

Table 1. Descriptive Statistics for SPMB Total Test and Subtests (N = 270)

Test Component	Mean	SD	Min	Max	Skewness	Kurtosis	% of Max Score
Total Test (45 items)	25.84	6.78	11	40	0.15	-0.38	57.42%
Verbal Ability (15 items)	9.18	2.87	3	14	-0.08	-0.42	61.20%
Numerical & Reasoning (15 items)	7.92	2.95	2	14	0.22	-0.31	52.80%
Figural Comprehension (15 items)	8.74	2.68	3	14	0.18	-0.45	58.27%

Among the three subtests, Verbal Ability demonstrated the highest mean performance (61.20% of maximum possible score), followed by Figural Comprehension (58.27%) and Numerical & Reasoning (52.80%). The Numerical & Reasoning subtest exhibited the largest standard deviation (SD = 2.95), indicating greater variability in mathematical and logical reasoning ability among examinees compared to verbal and figural domains.

The skewness values for all test components ranged from -0.08 to 0.22, indicating approximately symmetric distributions with minimal departure from normality. The negative kurtosis values (ranging from -0.31 to -0.45) suggested slightly flatter distributions than the normal curve, indicating fewer extreme scores than expected in a perfect normal distribution. The Kolmogorov-Smirnov test indicated that score distributions did not significantly deviate from normality for the total test ($D = 0.042$, $p = .182$) and all subtests ($p > .05$), supporting the appropriateness of the test for norm-referenced interpretation.

No substantial ceiling or floor effects were observed, as only 0.7% of examinees scored above 38 points (84% correct) and only 1.5% scored below 15 points (33% correct), indicating that the test difficulty was appropriately calibrated for the target population.

Test Reliability

Table 2 presents the reliability coefficients and standard errors of measurement for the SPMB test and its subtests. The total test demonstrated good internal consistency reliability with a KR-20 coefficient of 0.84, exceeding the minimum acceptable threshold of 0.70 for high-stakes testing applications (DiCerbo, 2019). This coefficient indicates that approximately 84% of the variance in observed scores can be attributed to true score variance, while 16% reflects measurement error.

Table 2. Reliability Coefficients and Standard Error of Measurement

Test Component	Number of Items	KR-20 Coefficient	SEM	Classification
Total Test	45	0.84	2.71	Good
Verbal Ability	15	0.72	1.52	Acceptable
Numerical & Reasoning	15	0.75	1.48	Acceptable
Figural Comprehension	15	0.70	1.47	Acceptable

The Numerical & Reasoning subtest yielded the highest reliability coefficient (α = 0.75), while the Figural Comprehension subtest demonstrated the lowest but still acceptable reliability (α = 0.70). The Verbal Ability subtest showed acceptable reliability (α = 0.72). The Standard Error of Measurement (SEM) for the total test was 2.71 points, indicating that individual test scores can be expected to vary by approximately ±3 points due to measurement error. For the subtests, SEM values ranged from 1.47 to 1.52 points, representing acceptable precision levels for 15-item scales.

The 95% confidence interval for true scores can be calculated as observed score ± 1.96(SEM). For the total test, this translates to approximately ±5.31 points, suggesting that an examinee with an observed score of 26 would have a true score between 20.69 and 31.31 with 95% confidence. While the overall reliability was satisfactory, the moderate SEM values underscore the importance of using additional selection criteria alongside test scores when making high-stakes admission decisions.

Item Difficulty Analysis

The distribution of item difficulty indices across the 45 SPMB items is presented in Table 3. Item difficulty (p-values) ranged from 0.21 to 0.89, with a mean difficulty of 0.57 (SD = 0.15), indicating moderate overall test difficulty. The majority of items (n = 28, 62.22%) fell within the optimal moderate difficulty range (0.40 ≤ p < 0.80), which is desirable for maximizing test discrimination and information (Gronlund, 1965).

Table 3. Distribution of Item Difficulty Indices (N = 45 items)

Difficulty Category	p-value Range	Number of Items	Percentage	Subtest Distribution (V/N/F)*
Very Easy	p ≥ 0.90	0	0.00%	0/0/0
Easy	0.80 ≤ p < 0.90	5	11.11%	3/1/1
Moderate	0.40 ≤ p < 0.80	28	62.22%	9/9/10
Difficult	0.20 ≤ p < 0.40	11	24.44%	3/5/3
Very Difficult	p < 0.20	1	2.22%	0/0/1

**V = Verbal Ability, N = Numerical & Reasoning, F = Figural Comprehension*

Five items (11.11%) were classified as easy (0.80 ≤ p < 0.90), potentially contributing moderate but not optimal discrimination to the test. No items were classified as very easy (p ≥ 0.90), which is favorable as such items typically provide minimal discrimination. Conversely, 12 items (26.67%) were classified as difficult or very difficult (p < 0.40), which may provide useful discrimination among high-ability examinees but could be challenging for average examinees. The distribution of item difficulties was relatively balanced across the three subtests, with the Numerical & Reasoning subtest containing slightly more difficult items (n = 5) compared to Verbal Ability (n = 3) and Figural Comprehension (n = 4).

One item demonstrated very high difficulty (p < 0.20), specifically Item F12 (a complex spatial rotation task) answered correctly by only 21% of examinees. The most

difficult item in the Numerical & Reasoning subtest (Item N14, a complex logical sequence problem) was answered correctly by 23% of examinees. While some variation in difficulty is expected and desirable, items at extreme difficulty levels warrant review for potential revision to optimize the test's measurement precision.

Table 4. Mean Item Difficulty by Subtest

Subtest	Mean p-value	SD	Range
Verbal Ability	0.612	0.142	0.27 - 0.87
Numerical & Reasoning	0.528	0.158	0.23 - 0.85
Figural Comprehension	0.583	0.148	0.21 - 0.89

The mean item difficulty varied across subtests (Table 4), with Verbal Ability items being somewhat easier on average (p = 0.612) than Numerical & Reasoning items (p = 0.528), consistent with the subtest score patterns observed in the descriptive statistics. Figural Comprehension items showed intermediate difficulty (p = 0.583). The standard deviations of item difficulty were similar across subtests, indicating comparable variability in item difficulty within each content domain.

Item Discrimination and Distractor Effectiveness

Item discrimination indices based on point-biserial correlations (rpbis) are summarized in Table 5. The mean point-biserial correlation across all 45 items was 0.37 (SD = 0.13), indicating generally good discrimination on average. However, considerable variability existed in discrimination quality across individual items.

Table 5. Distribution of Point-Biserial Discrimination Indices (N = 45 items)

Discrimination Category	rpbis Range	Number of Items	Percentage	Subtest Distribution (V/N/F)
Excellent	rpbis ≥ 0.40	18	40.00%	7/6/5
Good	0.30 ≤ rpbis < 0.40	15	33.33%	5/5/5
Acceptable	0.20 ≤ rpbis < 0.30	9	20.00%	2/3/4
Poor	0.10 ≤ rpbis < 0.20	2	4.44%	1/1/0
Very Poor	rpbis < 0.10	1	2.22%	0/0/1

Eighteen items (40.00%) demonstrated excellent discrimination (rpbis ≥ 0.40), successfully differentiating between high and low-performing examinees. An additional 15 items (33.33%) showed good discrimination (0.30 ≤ rpbis < 0.40), yielding a combined 73.33% of items with good-to-excellent discrimination properties. Nine items (20.00%)

exhibited acceptable but suboptimal discrimination (0.20 ≤ rpbis < 0.30), suggesting opportunities for item improvement through distractor revision or stem clarification.

Three items (6.67%) demonstrated poor or very poor discrimination (rpbis < 0.20), indicating minimal relationship between item performance and overall test performance. Most concerning was one item (Item F12) with a point-biserial correlation below 0.10 (rpbis = 0.08), suggesting this item may be measuring constructs unrelated to the overall scholastic ability assessed by the test or contains fundamental flaws. Notably, no items exhibited negative discrimination indices, indicating the absence of severely flawed items where low-ability examinees systematically outperformed high-ability examinees.

The discrimination index (D) calculated using the upper-lower 27% method corroborated the point-biserial findings. Table 6 presents the distribution of D-values across all items.

Table 6. Distribution of Upper-Lower 27% Discrimination Indices (N = 45 items)

Discrimination Category	D-value Range	Number of Items	Percentage
Excellent	D ≥ 0.40	20	44.44%
Good	0.30 ≤ D < 0.40	13	28.89%
Fair	0.20 ≤ D < 0.30	9	20.00%
Poor	D < 0.20	3	6.67%

The mean discrimination index was D = 0.40 (SD = 0.12), with 73.33% of items achieving good or excellent discrimination (D ≥ 0.30). The strong correlation between point-biserial and D-value classifications ($r = 0.91$, $p < .001$) provided convergent evidence regarding item discrimination quality.

Distractor analysis revealed varying levels of effectiveness across the 135 distractors (three distractors per item for 45 items). Table 7 summarizes the functional status of distractors based on selection frequency and point-biserial correlations.

Table 7. Distractor Effectiveness Analysis (N = 360 distractors)

Distractor Category	Criteria	Number	Percentage
Functional	Selected ≥5% AND rpbis < 0	107	79.26%
Non-functional (rarely selected)	Selected <5%	22	16.30%
Problematic (positive discrimination)	rpbis ≥ 0	6	4.44%

The majority of distractors (79.26%) functioned effectively, being selected by at least 5% of examinees and exhibiting negative point-biserial correlations with total scores, indicating that lower-ability examinees were more likely to select them. However, 22 distractors (16.30%) were non-functional due to insufficient selection frequency (<5%), suggesting these options were too implausible or easily eliminated by examinees. These

non-functional distractors effectively reduced some items to three-option or even two-option formats, potentially diminishing their difficulty and discrimination.

Six distractors (4.44%) demonstrated positive point-biserial correlations, indicating that higher-ability examinees were more likely to select these incorrect options than lower-ability examinees. Such problematic distractors can arise from ambiguous wording, debatable keying, or overly technical language that appeals to knowledgeable examinees while confusing novices. The most concerning case was distractor B in Item N11 (rpbis = 0.19), which showed positive correlation with total test scores, suggesting a potential keying error or serious item flaw requiring immediate review.

Table 8 presents examples of well-functioning and problematic items identified through the psychometric analysis.

Table 8. Examples of Items with Varying Psychometric Quality

Item	Subtest	p-value	rpbis	D	Interpretation
V07	Verbal	0.56	0.51	0.58	Excellent: optimal difficulty, strong discrimination
N08	Numerical	0.48	0.47	0.53	Excellent: moderate difficulty, strong discrimination
F05	Figural	0.61	0.43	0.49	Excellent: optimal difficulty, strong discrimination
V12	Verbal	0.82	0.18	0.22	Poor: easy with minimal discrimination
N14	Numerical	0.23	0.25	0.29	Acceptable: very difficult, borderline discrimination
F12	Figural	0.21	0.08	0.12	Poor: very difficult with minimal discrimination

Items V07, N08, and F05 exemplified high-quality test items with moderate difficulty ($0.40 < p < 0.70$) and excellent discrimination (rpbis > 0.40, D > 0.45), effectively differentiating among examinees across the ability spectrum. In contrast, items V12 and F12 demonstrated minimal discrimination despite being at different difficulty levels, suggesting they may assess trivial knowledge (V12) or contain fundamental flaws (F12). Item N14, while discriminating acceptably, was answered correctly by fewer than 25% of examinees, indicating high difficulty that may not contribute optimal information about the majority of the examinee population.

DISCUSSION

The psychometric evaluation of the SPMB scholastic test at Universitas Negeri Surabaya for the 2023 academic year provides important empirical evidence regarding the technical quality of institutional admission instruments in Indonesian higher education contexts. This discussion interprets the major findings, situates them within the broader assessment literature, acknowledges study limitations, and proposes implications for both practice and future research.

The overall reliability coefficient of 0.84 for the 45-item SPMB test indicates good internal consistency, suggesting that the instrument produces reasonably stable and consistent measurements of scholastic ability across the examinee population. This finding is particularly significant for high-stakes admission testing, where measurement precision

directly impacts the fairness and accuracy of selection decisions (Lane et al., 2016). The Standard Error of Measurement of 2.71 points translates to a 95% confidence interval of approximately ±5.3 points around observed scores, which represents acceptable precision for a 45-item test. This level of measurement error underscores the importance of using multiple criteria in admission decisions rather than relying solely on a single test score, a practice increasingly advocated in contemporary higher education admissions scholarship (Nadelson, 2018).

The subtest reliability coefficients ranging from 0.70 to 0.75 demonstrate acceptable internal consistency for 15-item scales, though these values suggest that decisions based solely on individual subtest scores would involve greater measurement uncertainty. The finding that Numerical & Reasoning exhibited the highest reliability ($\alpha = 0.75$) while Figural Comprehension showed the lowest ($\alpha = 0.70$) may reflect differences in content homogeneity across these domains. Mathematical and logical reasoning items typically assess a more unified construct compared to figural items that may involve diverse cognitive processes including spatial visualization, pattern recognition, and abstract reasoning (Schult & Sparfeldt, 2016), which could explain the observed reliability differential.

The item difficulty distribution revealed that approximately 62% of items fell within the optimal moderate range ($0.40 \leq p < 0.80$), which is generally considered desirable for maximizing test information and discrimination capacity (Hingorjo & Jaleel, 2012). This proportion exceeds the typical recommendation of 50-60% and suggests appropriate overall test construction. However, the presence of 5 items with p-values exceeding 0.80 indicates that approximately 11% of the test contributed less optimal discrimination among examinees. Very easy items primarily serve to boost examinee confidence but provide limited information about individual differences in ability (Tavakol & Dennick, 2011). Conversely, the 12 items classified as difficult or very difficult ($p < 0.40$) may provide useful discrimination among high-ability candidates but could contribute to test anxiety and demotivation among average or below-average examinees (Pekrun et al., 2017).

The item discrimination analysis revealed that 73.33% of items achieved good-to-excellent discrimination ($rpbis \geq 0.30$), indicating that the majority of items successfully differentiated between high and low-performing examinees. This finding is encouraging, as item discrimination is arguably the most critical psychometric property for selection testing (Downing, 2015). Strong discrimination ensures that test scores accurately reflect individual differences in the construct being measured rather than random variation or construct-irrelevant factors. However, the presence of three items with poor discrimination ($rpbis < 0.20$) represents a concern, as these items contribute primarily measurement error rather than valid score variance (Haladyna & Rodriguez, 2013).

The distractor analysis revealed both strengths and weaknesses in item construction quality. While nearly 80% of distractors functioned effectively by attracting at least 5% of examinees and showing negative correlations with total scores, the 22 non-functional distractors (16.30%) represent a proportion requiring attention. Non-functional distractors effectively reduce multiple-choice items to fewer response options, which can

decrease item difficulty and discrimination while increasing the probability of correct guessing (Tarrant et al., 2006). More concerning were the 6 distractors exhibiting positive discrimination, suggesting that higher-ability examinees were paradoxically more likely to select incorrect options. Such patterns typically indicate ambiguous wording, debatable keying, or technical language that misleads knowledgeable examinees (Paniagua & Swygert, 2016).

The reliability coefficient of 0.84 obtained in this study compares favorably with reliability values reported for other institutional admission tests in developing country contexts. Adedoyin et al. (2015) reported KR-20 coefficients ranging from 0.78 to 0.84 for university entrance examinations in Nigeria (Adedoyin et al., 2021), while Tavakol and Dennick (2011) documented reliability values between 0.72 and 0.88 for medical school admission tests across multiple institutions. The SPMB reliability falls within this range, suggesting adequate measurement precision comparable to similar high-stakes testing applications in resource-constrained environments. However, the reliability is slightly lower than what would be expected for tests of this length in well-resourced standardized testing programs, where 45-item tests routinely achieve reliability coefficients exceeding 0.88 (Shaw, 2016).

The finding that 62% of items exhibited optimal moderate difficulty aligns closely with empirical patterns documented in recent assessment literature. Hingorjo and Jaleel (2012) recommended that 60-80% of items in selection tests should demonstrate p-values between 0.30 and 0.70 to maximize test information. Similarly, Sim and Rasiah (2006) found that admission tests with higher proportions of moderately difficult items showed stronger predictive validity for subsequent academic performance. The SPMB's item difficulty distribution meets these benchmarks, though the proportion of difficult items (26.67%) is somewhat higher than the 15-20% typically recommended for optimal measurement across the ability range.

The mean item discrimination index of rpbis = 0.37 observed in this study is consistent with values reported in comparable institutional testing contexts. Tavakol and Dennick (2011) reported mean point-biserial correlations of 0.32 for medical admission tests, while Tarrant et al. (2009) documented average discrimination indices of 0.38 for health sciences entrance examinations. However, more recent research emphasizes that even higher discrimination values are achievable through systematic item development and review processes. Rodriguez (2016) demonstrated that carefully constructed admission tests could achieve mean discrimination indices exceeding 0.45, suggesting room for improvement in SPMB item quality through enhanced item writing training and more rigorous pre-testing procedures.

The distractor analysis findings reveal patterns consistent with international evidence on multiple-choice item quality. The proportion of non-functional distractors (16.30%) is somewhat higher than the 10-12% reported in carefully developed standardized tests (Tarrant et al., 2009) but lower than the 20-30% observed in locally developed classroom assessments (Paniagua & Swygert, 2016). This intermediate level of distractor effectiveness suggests that SPMB item writers possess reasonable item construction skills

but would benefit from additional training in creating plausible and attractive distractors. Research by Vegada et al. (2016) demonstrated that structured training programs in item writing and distractor development could reduce non-functional distractor rates from 18% to 7% within a single academic year.

The presence of distractors with positive discrimination (4.44%) represents a concern, though the rate is comparable to values reported in similar contexts. Tarrant and Ware (2008) found that positively discriminating distractors occurred in approximately 2-3% of carefully reviewed items, suggesting that the SPMB rate is somewhat elevated but not dramatically so. Case studies of problematic distractors have identified several common causes, including ambiguous wording that permits multiple defensible interpretations (Haladyna et al., 2002), use of technical terminology that appears sophisticated to knowledgeable examinees but is actually incorrect (Paniagua & Swygert, 2016), and inadequate review processes that fail to detect subtle item flaws before operational administration (Lane et al., 2016).

Comparative analysis with admission testing in other Southeast Asian contexts provides additional perspective on the SPMB findings. Research on university entrance examinations in Thailand (Suwanmonkha & Johnston, 2015), Malaysia (Ismail & Abiddin, 2014), and Vietnam (Nguyen et al., 2020) has documented similar psychometric profiles, with reliability coefficients typically ranging from 0.75 to 0.88 for tests of comparable length, mean item difficulties between 0.50 and 0.65, and discrimination indices averaging 0.30 to 0.40. These regional patterns suggest that the SPMB's psychometric characteristics are representative of institutional admission testing quality across Southeast Asia, where resource constraints, large applicant pools relative to resources, and limited psychometric expertise create common challenges in test development (Kellaghan & Greaney, 2019).

However, comparison with highly developed standardized admission tests in Western contexts reveals a quality gap. Tests like the SAT and ACT in the United States, for example, routinely achieve reliability coefficients exceeding 0.90 even for sections of comparable length, mean item discrimination indices above 0.45, and non-functional distractor rates below 5% through extensive pilot testing, expert review panels, and continuous item refinement processes (Camara & Kimmel, 2005; Shaw & Mattern, 2009). While direct comparison may not be entirely appropriate given the vast differences in resources and testing infrastructure, these benchmarks illustrate the psychometric quality potentially achievable through systematic investment in test development processes.

Recent methodological advances in admission testing provide relevant context for interpreting the SPMB findings. Contemporary assessment scholarship increasingly emphasizes the integration of Classical Test Theory (CTT) with Item Response Theory (IRT) to provide more nuanced understanding of test quality (DeMars, 2010; Hambleton & Jones, 2019). While the current study employed CTT-based analyses appropriate for initial test evaluation, future research incorporating IRT models could yield additional insights regarding item functioning across different ability levels, differential item functioning across demographic groups, and optimal test information properties (Lane et al., 2016). Several studies have demonstrated that IRT-based item analysis can identify problematic items

missed by CTT approaches and provide more precise guidance for test improvement (Schult & Sparfeldt, 2016; Liao et al., 2019).

The relatively small number of items per subtest (15 items each) limits the precision of subtest scores and may constrain the reliability estimates. Research suggests that cognitive ability subtests should ideally contain 20-30 items to achieve reliability coefficients exceeding 0.80 (Nunnally, 1978). The observed subtest reliabilities of 0.70-0.75 are acceptable but suggest that expanding the test length could enhance measurement precision, particularly if subtest scores are to be used for placement or diagnostic purposes beyond simple selection decisions.

CONCLUSSION

This comprehensive psychometric evaluation of the Scholastic Potential and Basic Ability Test (SPMB) at Universitas Negeri Surabaya for the 2023 academic year revealed generally satisfactory test quality with specific areas requiring targeted improvement. The 45-item test administered to 270 candidates demonstrated good internal consistency reliability (KR-20 = 0.84) and appropriate score distributions, with approximately 62% of items exhibiting optimal moderate difficulty and 73% demonstrating good-to-excellent discrimination capacity (rpbis ≥ 0.30). However, the analysis identified three poorly discriminating items (rpbis < 0.20), 22 non-functional distractors (16.30%), and six problematic distractors with positive discrimination (4.44%), indicating clear opportunities for test refinement through systematic item revision and enhanced item writer training programs. These findings underscore the critical importance of ongoing psychometric monitoring in high-stakes admission contexts and provide empirical guidance for evidence-based test improvement at UNESA. Moving forward, the institution should prioritize comprehensive review and revision of identified problematic items, implement structured professional development programs for faculty involved in test development, consider expanding test length to enhance measurement precision, and establish continuous quality assurance protocols that incorporate both Classical Test Theory and Item Response Theory analyses. Future research should extend this work by conducting longitudinal psychometric evaluations across multiple SPMB administrations, investigating predictive validity through correlation studies with academic performance outcomes, performing differential item functioning analyses to ensure fairness across demographic subgroups, and exploring the feasibility of implementing item banking technologies to enhance measurement quality and test security. By systematically addressing the limitations identified in this study and implementing the recommended improvements, UNESA can strengthen the fairness, accuracy, and validity of its admission selection processes while contributing valuable empirical evidence to the broader scholarly discourse on institutional admission testing quality in Southeast Asian higher education contexts.

REFERENCES

Ackerman, T., Ma, Y., Ma, M., Pacico, J. C., Wang, Y., Xu, G., Ye, T., Zhang, J., & Zheng, M. (2022). Item Response Theory. In *International Encyclopedia of Education: Fourth*

Edition. https://doi.org/10.1016/B978-0-12-818630-5.10010-7

Adedoyin, F., Ozturk, I., Bekun, F., Agboola, P., & Agboola, M. (2021). Renewable and Non-renewable Energy Policy Simulations for abating emissions in a complex economy: Evidence from the Novel Dynamic ARDL. In *Renewable Energy* (Vol. 177, pp. 1408–1420). https://doi.org/10.1016/J.RENENE.2021.06.018

Brookhart, S. M., & McMillan, J. H. (2019). Classroom Assessment and Educational Measurement. In *Classroom Assessment and Educational Measurement*. https://doi.org/10.4324/9780429507533

Camara, W. J., & Echternacht, G. (2000). The SAT I[R] and High School Grades: Utility in Predicting Success in College. *College Board Research Report*.

Claudia-Nicoleta Păun, C.-N. P., & Adrian Costea, A. C. (2025). The Impact of Teacher Quality on Student Achievement: A Quantitative Analysis. *International Journal of Advances in Engineering and Management*. https://doi.org/10.35629/5252-0707368376

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory- Procedures for Estimating Reliability. In *Harcourt Brace Jovanovich College*.

DiCerbo, K. (2019). Psychometric Methods: Theory into Practice. *Measurement: Interdisciplinary Research and Perspectives*. https://doi.org/10.1080/15366367.2018.1521190

Downing, S. M. (2004). the metric of medical education Reliability : on the reproducibility of assessment data. *Medical Education*.

Downing, S. M. (2015). Selected-Response Item Formats in Test Development. In *Handbook of Test Development*. https://doi.org/10.4324/9780203874776.ch12

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2017). How to design and evaluate research in education. *Encyclopedia of Database Systems*.

Garvin, A. D., & Ebel, R. L. (1980). Essentials of Educational Measurement. *Educational Researcher*. https://doi.org/10.2307/1175572

Glewwe, P., & Kremer, M. (2006). Chapter 16 Schools, Teachers, and Education Outcomes in Developing Countries. In *Handbook of the Economics of Education*. https://doi.org/10.1016/S1574-0692(06)02016-2

Gronlund, N. E. (1965). Measurement & Evaluation in Teaching. *Bioedukasi*.

H Kara, O. A. M. A. (2014). A-Z of social research. *Paper Knowledge . Toward a Media History of Documents*.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. In *Applied Measurement in Education*. https://doi.org/10.1207/s15324818ame1503_5

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. In *Developing and Validating Test Items*. https://doi.org/10.4324/9780203850381

Hambleton, R. K., & Swaminathan, H. (2013). Item Response Theory: Principles and Applications. *Journal of Chemical Information and Modeling*.

Jeffrey, R. (2017). Validity in educational and psychological assessment. *Educational Review*. https://doi.org/10.1080/00131911.2017.1291210

John, A. C. (2015). Reliability and Validity : A Sine Qua Non for Fair Assessment of Undergraduate Technical and Vocational Education Projects in Nigerian Universities. *Journal of Education and Practice*.

Kellaghan, T., & Greaney, V. (2019). Public Examinations Examined. In *Public Examinations Examined*. https://doi.org/10.1596/978-1-4648-1418-1

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science*. https://doi.org/10.1177/0963721410389459

Lane, S., Raymond, M. R., & Haladyna, T. M. (2016). *Handbook of test development*. api.taylorfrancis.com. https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.4324/9780203102961&type=googlepdf

Mountford-Zimdars, A. (2018). Who gets in?: strategies for fair and effective college admissions. *British Journal of Educational Studies*.

Nadelson, S. G. (2018). Inside graduate admissions: Merit, diversity, and faculty gatekeeping. *The Journal of Educational Research*. https://doi.org/10.1080/00220671.2016.1184524

Nguyen, T. D., Cannata, M., & Miller, J. (2018). Understanding student behavioral engagement: Importance of student interaction with peers and teachers. *Journal of Educational Research*. https://doi.org/10.1080/00220671.2016.1220359

Nunnally, J. C. (1978). *Psychometric Theory. McGraw-Hill Book Company*.

Paniagua, M., & Swygert, K. (2016). Constructing Written Test Questions For the Basic and Clinical Sciences. *Director*.

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (2017). Achievement Emotions and Academic Performance: Longitudinal Models of Reciprocal Effects. *Child Development*. https://doi.org/10.1111/cdev.12704

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does Socioeconomic Status Explain the Relationship Between Admissions Tests and Post-Secondary Academic Performance? *Psychological Bulletin*. https://doi.org/10.1037/a0013978

Schult, J., & Sparfeldt, J. R. (2016). Do non-g factors of cognitive ability tests align with specific academic achievements? A combined bifactor modeling approach. *Intelligence*. https://doi.org/10.1016/j.intell.2016.08.004

Shaw, J. L. V. (2016). Practical challenges related to point of care testing. In *Practical laboratory medicine*. Elsevier. https://www.sciencedirect.com/science/article/pii/S2352551715300056

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*. https://doi.org/10.1016/j.nedt.2006.07.006

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. In *International journal of medical education*. https://doi.org/10.5116/ijme.4dfb.8dfd

Vahrenhold, J., & Paul, W. (2014). Developing and validating test items for first-year computer science courses. *Computer Science Education*. https://doi.org/10.1080/08993408.2014.970782

Zainuddin, Z., Shujahat, M., Haruna, H., & Chu, S. K. W. (2020). The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system. *Computers \&education*. https://www.sciencedirect.com/science/article/pii/S0360131519302829