International Journal of Educational Management and Development Studies
Volume 7 Issue 2 June 2026
DOI: https://doi.org/10.53378/ijemds.353353

Analysis of item difficulty, discrimination and reliability of the KNP freshmen admission test: Basis for test revision and enhancement

[1]Mharfe M. Micaroz & [2]Jemuelle P. Frando

Abstract

Admission examinations support selection, profiling, and academic–support decisions in higher education; therefore, their score use must be supported by empirical evidence. This study evaluated the psychometric quality of the KNP Freshmen Admission Test by analyzing item difficulty, item discrimination, corrected item–total correlations, and internal consistency reliability. Using a descriptive–evaluative design, the study examined the dichotomously scored responses of 573 examinees to a 100–item admission test within the framework of Classical Test Theory. The test showed good internal consistency reliability (Cronbach's alpha = .824), but item–level results indicated a difficult test form and uneven item functioning: 65% of the items were difficult or very difficult, 58% showed poor or negative discrimination, and the integrated item decision framework retained 28 items, revised 31 items, and rejected 41 items. These findings indicate that the instrument has a reliable score structure but requires systematic item redevelopment, review of content alignment, and continuing validation before broader score–based admission decisions are made. The study contributes a practical institutional model for evidence–based admission test revision, item banking, and continuous assessment quality assurance.

Keywords: *admission testing, item analysis, educational measurement, institutional research*

Article History:

Received: *March 17, 2026* ***Revised***: *April 29, 2026*
Accepted: *April 30, 2026* ***Published online***: *May 19, 2026*

Suggested Citation:
Micaroz, M.M. & Frando, J.P. (2026). Analysis of item difficulty, discrimination and reliability of the KNP freshmen admission test: Basis for test revision and enhancement. *International Journal of Educational Management and Development Studies*, 7(2), 25-47. https://doi.org/10.53378/ijemds.353353

About the authors:
[1]Corresponding author. Doctor of Philosophy in Education major in Educational Leadership. Vice President for Academic Affairs, PAFTE/MTAP. Email: mharfemicaroz@gmail.com
[2]Master of Education major in Guidance and Counseling, Registered Guidance Counselor, Guidance Counselor

1. Introduction

Admission testing provides higher education institutions with a structured basis for screening applicants, profiling academic readiness, and identifying students who may need entry–level support. Because admission scores may influence consequential decisions, the quality of the test must be examined rather than assumed. Educational measurement standards emphasize that assessments used for decision–making should be supported by evidence on reliability, validity, fairness, and intended score use (American Educational Research Association [AERA], 2014; Crocker & Algina, 2008).

For locally developed admission tests, this responsibility is especially important. Local instruments are useful because they can reflect institutional priorities and applicant characteristics, but they may remain in use without repeated calibration, validity review, or item–level monitoring. Without such review, poorly functioning items may distort total scores, reduce fairness, and weaken the defensibility of selection decisions. Recent studies show that admission testing can contribute useful information when its technical properties are known and when results are interpreted within a broader transition–to–college framework (O'Neill & Nielsen, 2024; Watrin et al., 2022). At the same time, current discussions on admissions highlight fairness, bias, access, and the consequences of score use (Woo et al., 2023). In this study, these concerns are relevant because an overly difficult test or poorly discriminating items may affect how applicants are ranked and how institutional support needs are identified.

More recent work also shows that admission instruments should be evaluated beyond administrative convenience because their scores may be used for ranking, access decisions, academic advising, and early intervention. Predictive–validity studies in medical and higher education contexts indicate that admission measures can contribute to later academic–performance prediction when the instrument is technically sound and when interpretation is supported by outcome evidence (Almarabheh et al., 2022; Jaehn et al., 2025; Loder, 2024; Robb et al., 2025). Similarly, studies on computer–based and paper–based admission testing emphasize the need to examine whether test formats and test conditions preserve measurement equivalence before scores are used for consequential decisions (Escher et al., 2023).

Another issue concerns the construction and monitoring of multiple–choice admission items. Recent item–analysis studies show that item difficulty, discrimination, distractor effectiveness, and internal consistency are useful quality–control indicators for identifying items that should be retained, revised, or rejected (Bhattacherjee et al., 2022; Eleragi et al.,

2025; Ganji et al., 2025; Gebremichael et al., 2025; Kumar et al., 2021). These studies support the view that a local admission test should be treated as a living assessment system rather than as a fixed instrument that remains valid without recurring evidence.

The present study addresses a specific institutional gap: the KNP Freshmen Admission Test had not recently undergone a comprehensive psychometric evaluation using actual examinee response data. Thus, the institution lacked current evidence on whether the item pool was properly targeted, whether items differentiated stronger and weaker examinees, and whether the scores were internally consistent enough to support admission–related interpretation. Item analysis offers a practical starting point for addressing this gap. By examining difficulty, discrimination, corrected item–total correlations, and reliability, institutional researchers can identify which items may be retained, revised, or rejected. This process also supports item banking and future validation work, including studies on predictive validity and fairness across applicant groups.

Accordingly, this study analyzed the KNP Freshmen Admission Test using the responses of 573 examinees to a 100–item test form. The study was designed not to claim universal validity for the instrument but to generate local evidence for test revision, item redevelopment, and institutional assessment quality assurance. Specifically, the study sought to: describe the overall score profile of examinees in the KNP Freshmen Admission Test; determine the distribution of item difficulty across the 100–item test form; examine the discrimination power of each item using the upper–lower group method; estimate the internal consistency reliability of the test using Cronbach's alpha and item–total statistics; and classify items for retention, revision, or rejection as a basis for test enhancement.

2. Literature Review

2.1. Admission Testing and Educational Measurement

Admission examinations have historically been used to support student selection in situations where institutions require systematic evidence of readiness beyond prior grades or credentials. Although admission systems vary across countries and institutions, the common assumption underlying these examinations is that performance on the test can provide useful information about an applicant’s preparedness for the demands of higher education. Educational measurement scholars explain that tests used for selection should be judged not only on their administrative utility but also on their technical adequacy, including evidence

about item functioning, reliability, and score interpretation (Crocker & Algina, 2008; Nitko & Brookhart, 2014). In other words, the usefulness of an admission test depends not simply on its presence in the admission process, but on the quality of the data it produces. This expectation is echoed in the Standards for Educational and Psychological Testing, which emphasize that assessments used in consequential decision–making should be supported by evidence regarding reliability, validity, fairness, and intended use (AERA, 2014). In admission contexts, this requirement is especially important because test scores may affect acceptance, rejection, placement, or access to institutional opportunities. Consequently, when institutions employ admission examinations, they assume an obligation to review whether the instrument performs as expected in the population to which it is applied. These studies indicate that admission testing should be treated as both a measurement issue and a governance issue. Foundational standards emphasize reliability, validity, and fairness, while recent empirical studies show that test scores are useful only when institutions understand how items function and how scores relate to later student outcomes.

Recent studies further show that selection instruments should be evaluated through several forms of evidence. Watrin et al. (2022) reported that a psychology admission test could provide useful evidence when developed and validated for the intended student population, while Levacher et al. (2023) emphasized that admission tests should be examined for construct validity rather than assumed to measure the same academic–readiness construct. Escher et al. (2023) also showed that equivalence across delivery formats is important when institutions shift from paper–based to computer–based testing. These findings reinforce the need to examine not only whether an admission test produces scores, but whether the scores remain interpretable across test forms, cohorts, and administrative conditions.

In addition, recent studies on predictive validity and selection fairness extend the relevance of admission testing beyond initial screening. O'Neill and Nielsen (2024) connected admission testing with pre–academic exam self–efficacy and retention, while Jaehn et al. (2025), Loder (2024), and Robb et al. (2025) demonstrated that admission–related measures can be evaluated against later academic or selection outcomes. Woo et al. (2023) further argued that validity, fairness, and bias should be considered together when admissions evidence is used for consequential decisions. For the present study, this means that local item analysis is an initial but necessary step toward a broader validation program.

2.2. Item Difficulty and Test Targeting

Item difficulty is one of the most basic yet most informative indices in test analysis. In Classical Test Theory, item difficulty is usually defined as the proportion of examinees who answer an item correctly. Therefore, higher values indicate easier items, while lower values indicate more difficult items (Crocker & Algina, 2008). Ebel and Frisbie (1991) note that difficulty should not be understood as an inherently positive or negative trait. Rather, it must be interpreted in light of the purpose of the test, the characteristics of the examinee group, and the intended use of scores.

For selection and admission purposes, moderately difficult items are often preferred because they provide more information about variation in ability among applicants. If an item is too easy, a large proportion of examinees answer it correctly, which limits its usefulness in distinguishing stronger candidates from weaker ones. Conversely, if an item is too difficult, most examinees answer it incorrectly, which also reduces its value for differentiation. Thus, the practical issue is not whether an item is difficult in isolation, but whether the set of items contains an appropriate distribution of difficulty levels for the target population. The literature therefore converges on a central point: difficulty is not judged by severity alone but by fit between the item pool and the examinee population. Studies differ in context, but they agree that balanced difficulty strengthens interpretation and item banking. In the present study, this principle is applied to determine whether the KNP test is appropriately targeted to incoming applicants.

Recent studies confirm that difficulty indices remain useful for determining whether a test is appropriately matched to the examinee group. Bhattacherjee et al. (2022) and Kumar et al. (2021) showed that difficulty values help examiners identify items that may be too easy or too difficult for meaningful score interpretation. Iñarrairaegui et al. (2022) further demonstrated that the discriminative quality of MCQs may vary depending on students' academic level, indicating that difficulty must be interpreted in relation to the intended examinee population rather than in isolation. More recent evidence also links difficulty with item banking and outcome–based assessment. Ganji et al. (2025) examined the relationship between psychometric indices and course learning outcomes, while Eleragi et al. (2025) and Gebremichael et al. (2025) showed that difficulty, discrimination, reliability, and distractor indices can be used together to determine whether items should be stored, revised, or discarded.

These findings support the decision framework used in the present study, where items were not judged by difficulty alone but by their combined psychometric performance.

2.3. Item Discrimination and Item Functioning

Beyond difficulty, an effective test item must also discriminate. Item discrimination refers to the extent to which an item distinguishes between examinees who perform well on the overall test and those who perform poorly. Nitko and Brookhart (2014) explain that strong items tend to be answered correctly more often by high–scoring examinees than by low–scoring examinees. One common estimate of this relationship is the upper–lower group discrimination index, which compares the proportion of correct responses in a higher–performing group and a lower–performing group.

The practical value of discrimination analysis is substantial. An item may have a reasonable level of difficulty and still fail to discriminate if both stronger and weaker examinees respond similarly. In such cases, the item contributes little to ranking, screening, or differentiating examinees. More seriously, an item with negative discrimination suggests that lower–performing examinees were more likely to answer it correctly than higher–performing examinees. This pattern is generally regarded as a warning sign because it may indicate an incorrect answer key, ambiguous wording, construct–irrelevant difficulty, or other flaws in item construction (Ebel & Frisbie, 1991). Discrimination evidence complements difficulty evidence because it shows whether items contribute to the comparative purpose of admission testing. Theoretical sources agree that negative or near–zero discrimination signals possible item flaws, while recent applied studies show that such items often require key verification, distractor review, or replacement. This combined view guided the interpretation of weak and defective items in the present analysis.

Recent studies consistently identify discrimination as one of the strongest bases for revision decisions. Shahat (2024) and Mustafa and Hamid (2026) found that MCQ examinations may contain items with weak or suboptimal discrimination even when the test as a whole remains usable. Srisomsak et al. (2026) further showed that difficulty and discrimination indices can help detect flawed items, although fixed thresholds may miss some problematic questions. These studies suggest that negative or poor discrimination should trigger answer–key checking, content review, and distractor analysis.

The importance of discrimination is also visible outside medical education. Yüksel and Doğan (2022) reported that teacher–developed multiple–choice items may require correction when psychometric properties are insufficient, while Marsevani (2022) used difficulty, discrimination, and distractor efficiency to evaluate MCQs in language assessment. These studies support the interpretation used in the present study: items with poor or negative discrimination weaken the comparative purpose of admission testing and should not be reused without systematic review.

2.4. Reliability, Internal Consistency, and Institutional Test Quality

Reliability refers to the consistency of scores produced by an assessment instrument. In Classical Test Theory, reliability is concerned with the extent to which observed scores are free from random measurement error (Lord & Novick, 2008). For multi–item tests intended to measure a related domain, internal consistency reliability is commonly estimated using Cronbach's alpha. Tavakol and Dennick (2011) note that alpha provides a useful indication of the extent to which the items work together as a set. At the same time, recent reviews have cautioned that alpha should be interpreted carefully and not as a stand–alone index of test quality (Dorsah, 2026; Edelsbrunner et al., 2025).

A satisfactory alpha coefficient does not automatically mean that all items are functioning well. A test may yield an acceptable reliability coefficient while still containing several problematic items, especially if the test is long enough for the effects of weak items to be diluted by the stronger ones. This is one reason why reliability analysis is most informative when interpreted alongside item difficulty, item discrimination, and corrected item–total correlations. Reliability speaks to the consistency of the whole test, while item analysis speaks to the quality of its parts. Thus, reliability was interpreted as whole–test evidence, not as proof that all items were adequate. This distinction is important for institutional decision–making: a test may be reliable enough to produce consistent scores but still contain items that weaken fairness, efficiency, or content representation. The present study therefore combined alpha with item–level indices to avoid relying on a single statistic.

Recent research reinforces this cautious interpretation. Dorsah (2026) reviewed the use of Cronbach's alpha in educational research and emphasized that alpha is frequently misinterpreted when treated as a complete measure of instrument quality. Edelsbrunner et al. (2025) similarly showed that alpha in domain–specific knowledge tests is affected by test

characteristics and learning conditions. Therefore, alpha should be interpreted with item–level evidence, especially in an admission test where weak items may remain hidden within a long test form.

Contemporary studies also demonstrate that acceptable reliability can coexist with items needing revision. Salih et al. (2020), Ganji et al. (2025), Gebremichael et al. (2025), and Mustafa and Hamid (2026) used reliability together with item difficulty, discrimination, and distractor indices to evaluate MCQ quality. Their findings support the present study's combined interpretation of Cronbach's alpha, corrected item–total correlations, and item–level classification.

2.5. Theoretical framework

This study is anchored in Classical Test Theory (CTT), one of the most widely used frameworks in educational measurement. It assumes that an examinee's observed score is composed of a true score and an error score, often represented as Observed Score = True Score + Error. In this framework, test scores are not treated as perfect reflections of examinee ability. Instead, they are viewed as estimates influenced both by the examinee's actual level of competence and by imperfections in the measurement process (Crocker & Algina, 2008; Nitko & Brookhart, 2014). This basic premise makes CTT particularly useful for evaluating the functioning of a local admission examination.

Within this framework, item difficulty serves as an indicator of how accessible an item is to the examinee group, while item discrimination indicates whether the item aligns with the overall score pattern in a meaningful way. If stronger examinees tend to answer an item correctly and weaker examinees tend to answer it incorrectly, the item supports the measurement objective of the test. If this pattern is absent or reversed, the item may introduce noise into the score system. Consequently, difficulty and discrimination can be understood as practical quality–control indicators within CTT. Recent comparative work has shown that while Item Response Theory can offer richer parameter estimation, CTT remains highly useful for operational item review because its indicators are direct, transparent, and easily interpretable by institutional stakeholders (Ayanwale et al., 2022; Polat, 2022).

Furthermore, internal consistency reliability reflects the degree to which items act together as a coherent set. Cronbach's alpha, as used in this study, estimates whether responses across items show enough consistency to justify interpreting the test as a unified instrument.

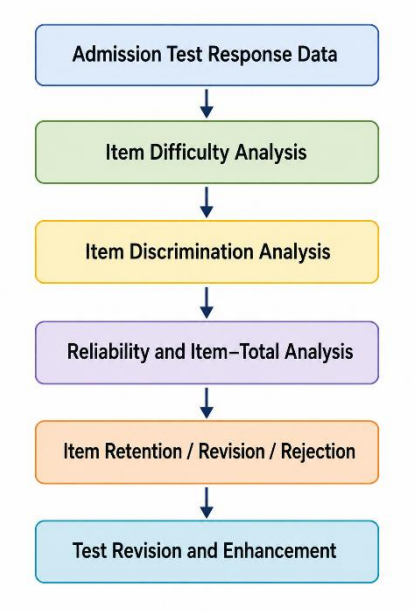
A higher alpha suggests that the items are sufficiently interrelated, although this does not eliminate the need for item–level review. In this sense, CTT provides a layered framework: the test as a whole is reviewed through reliability, while its individual components are reviewed through difficulty, discrimination, and item–total statistics.

This framework is appropriate because the KNP Freshmen Admission Test is locally developed, dichotomously scored, and intended for operational item review. CTT provides transparent indicators that can be readily interpreted by institutional stakeholders. However, because CTT indices are sample dependent, the results are interpreted as evidence for the 573–examinee cohort rather than as permanent item parameters. This limitation supports the need for repeated post–administration analysis and future studies using additional validity evidence.

Although CTT is practical, current literature also recognizes that its indices are sample dependent. Ayanwale et al. (2022) compared CTT and Item Response Theory (IRT) and noted that CTT remains useful because its indicators are transparent and accessible for institutional decision–making. Polat (2022) similarly showed that CTT and IRT can produce different performance information, suggesting that local institutions may begin with CTT while later extending validation through more advanced models. For this reason, the present study treats difficulty, discrimination, and reliability results as evidence for the specific KNP cohort and test form analyzed.

Figure 1

Conceptual framework of the study



Source: Adapted from Classical Test Theory concepts discussed by Crocker and Algina (2008) and Nitko and Brookhart (2014).

3. Methodology

This study employed a descriptive–evaluative research design. The design was appropriate because the purpose was to describe the psychometric performance of an existing admission test and evaluate item quality using established indicators. Rather than testing an intervention, the study generated evidence for revision and enhancement of an institutional screening instrument.

The dataset consisted of the responses of 573 examinees who completed the 100–item KNP Freshmen Admission Test. Each item was scored dichotomously, with 1 indicating a correct response and 0 indicating an incorrect response. The response matrix was reviewed for coding consistency and completeness before analysis. Because the study used actual institutional data, the findings are strongest for the examinee cohort and test form analyzed.

For item difficulty, the study used the proportion of examinees who answered each item correctly. Because the items were dichotomously scored, this value is equivalent to the item mean. Difficulty was interpreted using the following ranges: 0.00–0.20 as very difficult, 0.21–0.40 as difficult, 0.41–0.60 as moderately difficult, 0.61–0.80 as easy, and 0.81–1.00 as very easy. These ranges were adopted to provide a practical summary of the accessibility of the item pool and to help determine whether the test was appropriately targeted to the examinee population.

For item discrimination, the study used the upper–lower group method. Specifically, the discrimination index for each item was computed as the difference between the proportion of correct responses in the upper group and the proportion of correct responses in the lower group. Because the responses were dichotomously scored, the group mean for each item represented the proportion of correct responses in that group. The resulting discrimination indices were interpreted as follows: less than 0.00 as negative or defective, 0.00–0.19 as poor, 0.20–0.29 as marginal, 0.30–0.39 as good, and 0.40 and above as very good. This method was selected because it is widely used in operational item review and yields an immediately interpretable estimate of whether an item differentiates stronger examinees from weaker ones.

To estimate internal consistency reliability, Cronbach's alpha was computed for the full 100–item test. In addition, corrected item–total correlations and alpha–if–item–deleted values were reviewed to assess the contribution of individual items to the scale as a whole. These statistics allowed the study to move beyond global reliability and examine whether specific items were weakening or strengthening the internal structure of the test. Corrected

item–total correlations were especially useful in the integrated item–classification procedure because they indicate whether each item aligns with the overall score pattern independently of the item itself.

To make the results actionable, the study used an integrated item decision framework. Items were retained when difficulty was within 0.20 to 0.75, discrimination was at least 0.30, and corrected item–total correlation was at least 0.15. Items were rejected when they showed negative discrimination, corrected item–total correlation below 0.10, or extreme difficulty below 0.10 or above 0.90. All remaining items were marked for revision. These cutoffs were selected because they reflect common CTT practice while allowing institutionally practical decisions for a local admission test.

The choice to use an integrated item decision framework is also supported by recent assessment–quality studies. Eleragi et al. (2025), Ganji et al. (2025), and Gebremichael et al. (2025) used combined indices to guide decisions about item storage, revision, and improvement, while Srisomsak et al. (2026) cautioned that psychometric flags should be supplemented by expert review. Thus, the present study interpreted statistical results as decision–support evidence rather than as automatic proof of item validity or invalidity.

The analysis was also interpreted with attention to the assumptions and limits of CTT. Item difficulty and discrimination may vary across samples, and dichotomous scoring does not capture partial knowledge or reasoning processes. Therefore, the results should be used for institutional test improvement rather than broad generalization. Ethical safeguards were observed by using institutional response data for quality–assurance purposes, reporting only aggregated results, and avoiding the disclosure of individual examinee identities.

4. Findings and Discussion

Table 1

Descriptive statistics of the KNP freshmen admission test

Variable	N	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
Total Test Score	573	17.00	57.00	35.62	10.35	–0.083	–1.542
Reliability (Cronbach's Alpha)	573	–	–	.824	–	–	–

Table 1 shows that the mean total score was 35.62 out of 100, indicating that the average examinee answered slightly more than one–third of the test correctly. The highest

observed score was 57 and the lowest was 17, suggesting that the test was difficult for the cohort and did not use the upper portion of the possible score range. This pattern may reflect a demanding test blueprint, content not fully aligned with applicant readiness, item–construction problems, or cohort–level differences in preparation. Despite the low mean, the standard deviation of 10.35 indicates meaningful score variation. The nearly symmetrical skewness value (–0.083) and flat kurtosis (–1.542) suggest that the test still differentiated examinees across the observed range, although the distribution was centered at a difficult level. This result anticipates the item–level findings in Table 2.

The reliability coefficient of .824 indicates good internal consistency for the 100–item test. However, this result should be interpreted with caution because a long test can produce acceptable alpha even when several items perform poorly. Consistent with CTT, whole–test reliability must therefore be read alongside difficulty, discrimination, and corrected item–total statistics. This finding is consistent with recent item–analysis research showing that whole–test reliability may remain acceptable even when several individual items need revision. For example, Eleragi et al. (2025) reported acceptable reliability across examinations while still emphasizing the need for continuous item analysis, and Gebremichael et al. (2025) found that reliability levels varied across departments even when the same broad assessment approach was used. In the present study, the acceptable alpha therefore supports the internal consistency of the total score, but it does not remove the need for item–level improvement.

Table 2

Distribution of item difficulty

Difficulty Level	Frequency	Percentage
Very Difficult	17	17.0
Difficult	48	48.0
Moderately Difficult	23	23.0
Easy	10	10.0
Very Easy	2	2.0
Total	**100**	**100.0**

Table 2 shows a clear difficulty imbalance: 48 items were difficult and 17 were very difficult, meaning that 65% of the test fell in low–accessibility categories. Only 23 items were moderately difficult, 10 were easy, and 2 were very easy. This distribution suggests that the

test was not optimally targeted to the examinee cohort. Several explanations may account for this pattern. The test may include content beyond the academic preparation of many applicants, or some items may have ambiguous wording, overly complex stems, weak distractors, or misalignment with the intended competencies. Because admission tests must support both selection and diagnostic interpretation, a heavy concentration of difficult items may reduce the usefulness of scores for identifying borderline readiness and support needs. The implication is not to make the test uniformly easier but to rebalance the item pool. Future forms should include a deliberate spread of easy, moderate, and difficult items aligned with a table of specifications. This would improve score interpretation, strengthen fairness, and make the test more useful for both screening and student–support planning.

The high proportion of difficult items should be viewed as a targeting issue rather than as a simple indication that the test is academically rigorous. Kumar et al. (2021), Bhattacherjee et al. (2022), and Iñarrairaegui et al. (2022) showed that difficulty values are most useful when interpreted together with the examinee group and assessment purpose. For an admission test, too many difficult items may compress scores at the lower range and reduce the instrument's usefulness for identifying borderline readiness. This supports the recommendation to rebalance future test forms through a clearer table of specifications and pilot item review.

Table 3

Distribution of item discrimination

Discrimination Level	Frequency	Percentage
Negative / Defective	8	8.0
Poor	50	50.0
Marginal	14	14.0
Good	9	9.0
Very Good	19	19.0
Total	100	100.0

Table 3 identifies item discrimination as the most urgent technical concern. Fifty items were poor discriminators and eight were negative or defective; therefore, 58% of the item pool did not strongly distinguish higher–performing examinees from lower–performing examinees. In an admission test, this weakens the comparative function of the instrument. The results also show that the test contains a usable core: 19 items were very good discriminators and nine

were good discriminators. These items can serve as anchors for a revised form. By contrast, items with negative discrimination should be checked first for possible answer–key errors, ambiguous wording, implausible distractors, or content that rewards guessing rather than the intended competency.

When interpreted with the difficulty results, the discrimination findings suggest that item weakness may come from both targeting and construction problems. Difficult items are not automatically defective, but they must still distinguish strong and weak examinees. Items that are difficult and poorly discriminating should receive priority review because they contribute little to fair score interpretation.

Recent studies suggest that these weak and negative items should receive priority review. Shahat (2024), Mustafa and Hamid (2026), and Srisomsak et al. (2026) associated weak discrimination with possible item–writing flaws, answer–key concerns, and insufficient distractor functioning. Yüksel and Doğan (2022) likewise showed that teacher–developed items may require correction when validity and reliability evidence is insufficient. In the KNP test, items with negative discrimination should therefore be checked first for miskeying, ambiguity, or content that may not match the intended competency.

Table 4

Item classification based on psychometric criteria

Classification	Frequency	Percentage
Retain	28	28.0
Revise	31	31.0
Reject	41	41.0
Total	100	100.0

Table 4 translates the psychometric evidence into institutional decisions. Only 28 items met the retention criteria, indicating that less than one–third of the item pool was ready for direct reuse. These retained items should be preserved as preliminary anchors for an institutional item bank.

The 31 items marked for revision represent salvageable content. These items should undergo content review, stem rewriting, distractor strengthening, and alignment checking against the table of specifications before reuse. Revision is appropriate when an item shows some promise but does not yet satisfy all psychometric criteria.

The 41 rejected items require replacement rather than simple editing. This proportion indicates that test enhancement should be systematic rather than cosmetic. A structured cycle of blueprinting, expert review, pilot testing, item analysis, and item banking is needed to prevent the repeated use of weak items.

Based on the results, the test is reliable as a whole but uneven at the item level. This distinction is important: reliability supports cautious score interpretation, whereas item analysis identifies which parts of the test strengthen or weaken that interpretation. The results therefore justify revision without requiring the institution to abandon the entire test.

For admission policy, the findings point to three immediate actions. First, retain the 28 strongest items as the starting pool for future forms. Second, revise the 31 salvageable items through expert and psychometric review. Third, replace the 41 rejected items with newly developed questions based on a clearer test blueprint. These actions would improve measurement efficiency and support more defensible admission decisions.

The findings also provide a baseline for future validity work. After revision, the institution may examine predictive validity by relating admission scores to first–year performance, retention, and placement outcomes. Fairness studies may also compare item functioning across applicant subgroups when data are available. These next steps would extend the present CTT evidence into a broader validation program and align with published work emphasizing rigorous instrument development and assessment models in higher education (Asirit, 2024; Schutte, 2024).

The classification results also support the development of an institutional item bank. Recent studies recommend that items with acceptable difficulty, strong discrimination, and functioning distractors be stored for reuse, whereas weak items should be revised or removed from the operational pool (Eleragi et al., 2025; Ganji et al., 2025; Gebremichael et al., 2025; Gottlieb et al., 2023). The KNP results therefore provide a practical starting inventory: retained items may serve as anchors, revised items may be improved through expert review, and rejected items should be replaced by newly written items aligned with the test blueprint.

The findings further indicate that future enhancement may consider both human and technology–supported item development, but only with expert validation. Rezigalla (2024) and Cheung et al. (2023) showed that AI tools can generate MCQs, yet they also highlight the need for quality review, psychometric checking, and expert judgment. Capan Melser et al. (2020) and Gottlieb et al. (2023) similarly emphasize that strong MCQs require deliberate

construction, cognitive alignment, plausible distractors, and review against intended learning outcomes. Thus, item redevelopment for the KNP admission test should not rely on item generation alone; it should include blueprinting, expert validation, pilot testing, and post–administration analysis.

5. Limitations and Future Directions

The study is limited to one institutional test form and one cohort of 573 examinees; therefore, the findings should not be generalized to other admission tests or applicant populations without additional evidence. The analysis also relied on Classical Test Theory, whose item indices are sample dependent, and on dichotomous scoring, which does not capture partial reasoning. Future research should replicate item analysis across succeeding administrations, conduct predictive–validity studies using first–year academic outcomes, and examine fairness or differential item functioning when subgroup data become available.

Future validation should also examine whether the revised admission test predicts relevant student outcomes. Recent studies have connected admission measures with later academic performance, retention, or selection outcomes (Almarabheh et al., 2022; Jaehn et al., 2025; Loder, 2024; O'Neill & Nielsen, 2024; Robb et al., 2025). In addition, fairness–oriented analyses should be considered when applicant subgroup data become available, because admissions evidence should be interpreted with attention to validity, bias, and score–use consequences (Woo et al., 2023).

6. Conclusion

This study examined the psychometric quality of the KNP Freshmen Admission Test through item difficulty, item discrimination, corrected item–total correlation, and internal consistency reliability. The test demonstrated good internal consistency (Cronbach's alpha = .824), but item–level evidence showed that the form was difficult and that many items had weak or defective discrimination. The integrated decision framework retained 28 items, marked 31 for revision, and rejected 41.

The practical conclusion is that the admission test should be systematically revised rather than used repeatedly in its present form. The institution may follow a step–by–step enhancement roadmap: (1) prepare a revised table of specifications, (2) retain the strongest

items as anchors, (3) revise salvageable items through expert review, (4) replace rejected items with newly written questions, (5) pilot the revised form, (6) conduct post–administration item analysis, and (7) build an item bank supported by recurring reliability and validity checks.

This roadmap links test revision to broader institutional policy. By treating admission testing as a continuous quality–assurance process, the institution can strengthen fairness, improve score interpretation, and support evidence–based decisions on admission, placement, and academic intervention.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was not supported by any funding.

References

Almarabheh, A., Ismaeel, A., Al–Qahtani, A., & Al–Mutairi, A. (2022). Predictive validity of admission criteria in predicting academic performance of medical students: A retrospective cohort study. *Advances in Medical Education and Practice, 13*, 1009–1019. https://doi.org/10.2147/AMEP.S376792

American Educational Research Association (2014). *Standards for educational and psychological testing*. https://www.aera.net/Portals/38/1999%20Standards_revised.pdf

Asirit, L. B. L. (2024). From insight to measurement: A self–assessment tool development for entry–level teachers' instructional competence. *International Journal of Educational Management and Development Studies, 5*(1), 27–53. https://doi.org/10.53378/353043

Ayanwale, M. A., Chere–Masopha, J., & Morena, M. C. (2022). The classical test or item response measurement theory: The status of the framework at the Examination Council of Lesotho. *International Journal of Learning, Teaching and Educational Research, 21*(8), 384–402. https://doi.org/10.26803/ijlter.21.8.22

Bhattacherjee, S., Mukherjee, A., Bhandari, K., & Rout, A. J. (2022). Evaluation of multiple–choice questions by item analysis, from an online internal assessment of 6th semester medical students in a rural medical college, West Bengal. *Indian Journal of Community Medicine, 47*(1), 92–95. https://doi.org/10.4103/ijcm.ijcm_1156_21

Capan Melser, M., Steiner–Hofbauer, V., Lilaj, B., Agis, H., Knaus, A., & Holzinger, A. (2020). Knowledge, application and how about competence? Qualitative assessment of multiple–choice questions for dental students. *Medical Education Online, 25*(1), Article 1714199. https://doi.org/10.1080/10872981.2020.1714199

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T.–H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions–A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLOS ONE, 18*(8), Article e0290691. https://doi.org/10.1371/journal.pone.0290691

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Dorsah, P. (2026). The use of Cronbach's alpha reliability in educational research: A systematic review. *European Journal of Contemporary Education and E–Learning, 4*(2), 39–50. https://doi.org/10.59324/ejceel.2026.4(2).04

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

Edelsbrunner, P. A., Simonsmeier, B. A., & Schneider, M. (2025). The Cronbach's alpha of domain–specific knowledge tests before and after learning: A meta–analysis of published studies. *Educational Psychology Review, 37*, Article 4. https://doi.org/10.1007/s10648–024–09982–y

Eleragi, A. M. S., Miskeen, E., Hussein, K., Rezigalla, A. A., Adam, M. I. E., Al–Faifi, J. A., Alhalafi, A., Al Ameer, A. Y., & Mohammed, O. A. (2025). Evaluating the multiple–choice questions quality at the College of Medicine, University of Bisha, Saudi Arabia: A three–year experience. *BMC Medical Education, 25*, Article 233. https://doi.org/10.1186/s12909–025–06700–2

Escher, M., Weppert, D., Amelung, D., Huelmann, T., Stegt, S., & Hissbach, J. (2023). Paper–based and computer–based admission tests for medicine–Are they equivalent? *Frontiers in Education, 8*, Article 1209212. https://doi.org/10.3389/feduc.2023.1209212

Ganji, K. K., Ananthakrishnan, N., Manivasakan, S., Alruwaili, M. K., Alonazi, M. A., & Algarni, H. A. (2025). Analyzing the relationship between psychometric indices of item analysis with attainment of course learning outcomes: Cross–sectional study in integrated outcome–based dental curriculum courses. *BMC Medical Education, 25*, Article 1366. https://doi.org/10.1186/s12909–025–07871–8

Gebremichael, M. W., Baraki, B., Mehari, M.–A., & Assalfew, B. (2025). Item analysis of multiple choice questions from assessment of health sciences students, Tigray, Ethiopia. *BMC Medical Education, 25*, Article 441. https://doi.org/10.1186/s12909–025–06904–6

Gottlieb, M., Bailitz, J., Fix, M., Shappell, E., & Wagner, M. J. (2023). Educator's blueprint: A how–to guide for developing high–quality multiple–choice questions. *AEM Education and Training, 7*(1), Article e10836. https://doi.org/10.1002/aet2.10836

Iñarrairaegui, M., Fernández–Ros, N., Lucena, F., Landecho, M. F., García, N., Quiroga, J., & Herrero, J. I. (2022). Evaluation of the quality of multiple–choice questions according to the students' academic level. *BMC Medical Education, 22*, Article 779. https://doi.org/10.1186/s12909–022–03844–3

Jaehn, M., Hissbach, J., Frickhoeffer, M., Weppert, D., Zimmerhofer, A., Wolf, K., & Hampe, W. (2025). Predictive validity of admission tests and educational attainment on preclinical academic performance: A multisite study. *BMC Medical Education, 25*, Article 1255. https://doi.org/10.1186/s12909–025–07974–2

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical*

Journal Armed Forces India, 77(Suppl. 1), S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.007

Levacher, J., Koch, M., Stegt, S. J., Hissbach, J., Spinath, F. M., Escher, M., & Becker, N. (2023). The construct validity of the main student selection tests for medical studies in Germany. *Frontiers in Education, 8*, Article 1120129. https://doi.org/10.3389/feduc.2023.1120129

Loder, A. K. F. (2024). Student performance correlates of psychology admission exam scores and the number of places for students. *Acta Psychologica, 250*, Article 104523. https://doi.org/10.1016/j.actpsy.2024.104523

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Information Age Publishing.

Marsevani, M. (2022). Item analysis of multiple–choice questions. *English Review: Journal of English Education, 10*(3), 759–766. https://doi.org/10.25134/erjee.v10i3.6241

Mustafa, S., & Hamid, O. E. (2026). Psychometric item/question analysis of multiple–choice questions in fixed prosthodontics exam. *BMC Medical Education, 26*, Article 86. https://doi.org/10.1186/s12909–025–08429–4

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (7th ed.). Pearson.

O'Neill, L. D., & Nielsen, T. (2024). Admission testing, pre–academic exam self–efficacy, and retention: A prospective cohort study. *Studies in Educational Evaluation, 83*, Article 101383. https://doi.org/10.1016/j.stueduc.2024.101383

Polat, M. (2022). Comparison of performance measures obtained from foreign language tests according to item response theory vs. classical test theory. *International Online Journal of Education and Teaching, 9*(1), 471–485.

Rezigalla, A. A. (2024). AI in medical education: Uses of AI in construction type A MCQs. *BMC Medical Education, 24*, Article 247. https://doi.org/10.1186/s12909–024–05250–3

Robb, C., Banks, P. W., Copeland, H. L., MacIntosh, A., Ivan, R., Moskowitz, J. B., Reiter, H., & Sitarenios, G. (2025). Examining the predictive validity of an open–response situational judgment test with typed–response and video–response items. *Educational Assessment*, 1–11. https://doi.org/10.1080/10627197.2025.2576228

Salih, K. E. M. A., Jibo, A. M., Ishaq, M., Khan, S., Mohammed, O. A., & Al–Shahrani, A. M. (2020). Psychometric analysis of multiple–choice questions in an innovative curriculum in Kingdom of Saudi Arabia. *Journal of Family Medicine and Primary Care, 9*(7), 3663–3668. https://doi.org/10.4103/jfmpc.jfmpc_1034_19

Schutte, F. (2024). A model for assessments in higher education institutions. *International Journal of Educational Management and Development Studies, 5*(3), 92–117. https://doi.org/10.53378/ijemds.353088

Shahat, K. A. (2024). Item analysis of multiple–choice question (MCQ)–based exam efficiency among postgraduate pediatric medical students: An observational, cross–sectional study from Saudi Arabia. *Cureus, 16*(9), Article e69151. https://doi.org/10.7759/cureus.69151

Srisomsak, V., Sitticharoon, C., Keadkraichaiwat, I., Meethes, S., & Inpaen, I. (2026). Detection of flawed multiple–choice questions in preclinical medical education using item difficulty and discrimination indices: A six–year analysis. *BMC Medical Education, 26*, Article 92. https://doi.org/10.1186/s12909–025–08204–5

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

Watrin, L., Geiger, M., Levacher, J., Spinath, B., & Wilhelm, O. (2022). Development and initial validation of an admission test for bachelor psychology studies. *Frontiers in Education, 7*, Article 909818. https://doi.org/10.3389/feduc.2022.909818

Woo, S. E., LeBreton, J. M., Keith, M. G., & Tay, L. (2023). Bias, fairness, and validity in graduate–school admissions: A psychometric perspective. *Perspectives on Psychological Science, 18*(1), 3–31. https://doi.org/10.1177/17456916211055374

Yüksel, K. B., & Doğan, N. (2022). Investigation of psychometric properties of multiple–choice items developed by Turkish teachers. *Sakarya University Journal of Education, 12*(1), 130–149. https://doi.org/10.19126/suje.1007897

Appendix A

Complete 100–item analysis table

Item	Difficulty	Difficulty Level	Upper Mean	Lower Mean	D	Discrimination	CITC	Alpha if Deleted	Decision
I1	0.5113	Moderately Difficult	0.6278	0.3672	0.2606	Marginal	0.238	0.822	Revise
I2	0.3159	Difficult	0.3659	0.2539	0.112	Poor	0.111	0.824	Revise
I3	0.4136	Moderately Difficult	0.5457	0.25	0.2957	Marginal	0.285	0.821	Revise
I4	0.3089	Difficult	0.3785	0.2227	0.1558	Poor	0.147	0.823	Revise
I5	0.3194	Difficult	0.388	0.2344	0.1536	Poor	0.123	0.824	Revise
I6	0.548	Moderately Difficult	0.7855	0.2539	0.5316	Very Good	0.507	0.817	Retain
I7	0.822	Very Easy	0.9243	0.6953	0.229	Marginal	0.269	0.821	Revise
I8	0.4921	Moderately Difficult	0.7192	0.2109	0.5083	Very Good	0.496	0.817	Retain
I9	0.5288	Moderately Difficult	0.735	0.2734	0.4616	Very Good	0.428	0.818	Retain
I10	0.4241	Moderately Difficult	0.612	0.1914	0.4206	Very Good	0.386	0.819	Retain
I11	0.3281	Difficult	0.4006	0.2383	0.1623	Poor	0.163	0.823	Revise
I12	0.3333	Difficult	0.4669	0.168	0.2989	Marginal	0.299	0.821	Revise
I13	0.7225	Easy	0.8297	0.5898	0.2399	Marginal	0.236	0.822	Revise
I14	0.897	Very Easy	0.9811	0.793	0.1881	Poor	0.284	0.822	Revise
I15	0.6771	Easy	0.8896	0.4141	0.4755	Very Good	0.479	0.817	Retain
I16	0.1012	Difficult	0.164	0.0234	0.1406	Poor	0.235	0.822	Revise
I17	0.089	Very Difficult	0.1073	0.0664	0.0409	Poor	0.032	0.824	Reject
I18	0.2443	Difficult	0.3659	0.0938	0.2721	Marginal	0.297	0.821	Revise
I19	0.3072	Difficult	0.388	0.207	0.181	Poor	0.178	0.823	Revise
I20	0.1955	Very Difficult	0.2492	0.1289	0.1203	Poor	0.121	0.823	Revise
I21	0.4991	Moderately Difficult	0.7287	0.2148	0.5139	Very Good	0.479	0.817	Retain
I22	0.2513	Difficult	0.3438	0.1367	0.2071	Marginal	0.221	0.822	Revise
I23	0.4712	Moderately Difficult	0.6814	0.2109	0.4705	Very Good	0.454	0.818	Retain
I24	0.2269	Difficult	0.2429	0.207	0.0359	Poor	0.004	0.825	Reject

Item	Difficulty	Difficulty Level	Upper Mean	Lower Mean	D	Discrimination	CITC	Alpha if Deleted	Decision
I25	0.26	Difficult	0.2555	0.2656	–0.0101	Negative	–0.061	0.826	Reject
I26	0.2007	Difficult	0.2177	0.1797	0.038	Poor	0.033	0.825	Reject
I27	0.2112	Difficult	0.1703	0.2617	–0.0914	Negative	–0.149	0.827	Reject
I28	0.3403	Difficult	0.3785	0.293	0.0855	Poor	0.044	0.825	Reject
I29	0.2112	Difficult	0.1924	0.2344	–0.0420	Negative	–0.074	0.826	Reject
I30	0.1187	Very Difficult	0.1167	0.1211	–0.0044	Negative	–0.017	0.825	Reject
I31	0.3857	Difficult	0.5931	0.1289	0.4642	Very Good	0.45	0.818	Retain
I32	0.281	Difficult	0.4322	0.0938	0.3384	Good	0.35	0.82	Retain
I33	0.2251	Difficult	0.3596	0.0586	0.301	Good	0.377	0.82	Retain
I34	0.5009	Moderately Difficult	0.7098	0.2422	0.4676	Very Good	0.43	0.818	Retain
I35	0.4398	Moderately Difficult	0.6151	0.2227	0.3924	Good	0.387	0.819	Retain
I36	0.1065	Very Difficult	0.1167	0.0937	0.023	Poor	0.021	0.824	Reject
I37	0.1291	Very Difficult	0.142	0.1133	0.0287	Poor	0.006	0.825	Reject
I38	0.7766	Easy	0.9022	0.6211	0.2811	Marginal	0.288	0.821	Revise
I39	0.1204	Very Difficult	0.0946	0.1523	–0.0577	Negative	–0.100	0.826	Reject
I40	0.5846	Moderately Difficult	0.8328	0.2773	0.5555	Very Good	0.522	0.816	Retain
I41	0.6684	Easy	0.8612	0.4297	0.4315	Very Good	0.428	0.818	Retain
I42	0.4188	Moderately Difficult	0.6088	0.1836	0.4252	Very Good	0.409	0.818	Retain
I43	0.5777	Moderately Difficult	0.7886	0.3164	0.4722	Very Good	0.445	0.818	Retain
I44	0.7417	Easy	0.9054	0.5391	0.3663	Good	0.374	0.819	Retain
I45	0.5375	Moderately Difficult	0.6909	0.3477	0.3432	Good	0.315	0.82	Retain
I46	0.4538	Moderately Difficult	0.5615	0.3203	0.2412	Marginal	0.215	0.822	Revise
I47	0.6387	Easy	0.8486	0.3789	0.4697	Very Good	0.469	0.817	Retain
I48	0.6265	Easy	0.8202	0.3867	0.4335	Very Good	0.418	0.818	Retain
I49	0.7051	Easy	0.9306	0.4258	0.5048	Very Good	0.518	0.817	Retain
I50	0.3682	Difficult	0.5584	0.1328	0.4256	Very Good	0.421	0.818	Retain
I51	0.4276	Moderately Difficult	0.5489	0.2773	0.2716	Marginal	0.245	0.821	Revise
I52	0.3159	Difficult	0.4006	0.2109	0.1897	Poor	0.162	0.823	Revise
I53	0.2216	Difficult	0.2397	0.1992	0.0405	Poor	–0.013	0.825	Reject
I54	0.1449	Very Difficult	0.164	0.1211	0.0429	Poor	0.04	0.824	Reject
I55	0.2391	Difficult	0.2461	0.2305	0.0156	Poor	–0.015	0.826	Reject
I56	0.2234	Difficult	0.2524	0.1875	0.0649	Poor	0.042	0.825	Reject
I57	0.3264	Difficult	0.3785	0.2617	0.1168	Poor	0.089	0.824	Reject
I58	0.2234	Difficult	0.2713	0.1641	0.1072	Poor	0.078	0.824	Reject
I59	0.1134	Very Difficult	0.1136	0.1133	0.0003	Poor	–0.029	0.825	Reject

Item	Difficulty	Difficulty Level	Upper Mean	Lower Mean	D	Discrimination	CITC	Alpha if Deleted	Decision
I60	0.3543	Difficult	0.4132	0.2812	0.132	Poor	0.108	0.824	Revise
I61	0.2565	Difficult	0.2681	0.2422	0.0259	Poor	–0.021	0.826	Reject
I62	0.2199	Difficult	0.265	0.1641	0.1009	Poor	0.101	0.824	Revise
I63	0.2914	Difficult	0.2934	0.2891	0.0043	Poor	–0.026	0.826	Reject
I64	0.3176	Difficult	0.3785	0.2422	0.1363	Poor	0.111	0.824	Revise
I65	0.1937	Very Difficult	0.2177	0.1641	0.0536	Poor	0.046	0.824	Reject
I66	0.1134	Very Difficult	0.1293	0.0937	0.0356	Poor	0.028	0.824	Reject
I67	0.1658	Very Difficult	0.205	0.1172	0.0878	Poor	0.09	0.824	Reject
I68	0.3682	Difficult	0.429	0.293	0.136	Poor	0.091	0.824	Reject
I69	0.4206	Moderately Difficult	0.489	0.3359	0.1531	Poor	0.119	0.824	Revise
I70	0.2531	Difficult	0.3028	0.1914	0.1114	Poor	0.095	0.824	Reject
I71	0.3455	Difficult	0.4069	0.2695	0.1374	Poor	0.097	0.824	Reject
I72	0.356	Difficult	0.4826	0.1992	0.2834	Marginal	0.258	0.821	Revise
I73	0.3421	Difficult	0.4322	0.2305	0.2017	Marginal	0.174	0.823	Revise
I74	0.4974	Moderately Difficult	0.6404	0.3203	0.3201	Good	0.295	0.821	Retain
I75	0.3839	Difficult	0.4858	0.2578	0.228	Marginal	0.205	0.822	Revise
I76	0.2199	Difficult	0.2524	0.1797	0.0727	Poor	0.045	0.825	Reject
I77	0.2112	Difficult	0.2429	0.1719	0.071	Poor	0.081	0.824	Reject
I78	0.2426	Difficult	0.2618	0.2188	0.043	Poor	0.016	0.825	Reject
I79	0.1832	Very Difficult	0.1956	0.168	0.0276	Poor	0.007	0.825	Reject
I80	0.1414	Very Difficult	0.1388	0.1445	–0.0057	Negative	–0.033	0.825	Reject
I81	0.6771	Easy	0.8423	0.4727	0.3696	Good	0.366	0.819	Retain
I82	0.4485	Moderately Difficult	0.6341	0.2188	0.4153	Very Good	0.386	0.819	Retain
I83	0.4346	Moderately Difficult	0.5836	0.25	0.3336	Good	0.318	0.82	Retain
I84	0.2723	Difficult	0.3438	0.1836	0.1602	Poor	0.148	0.823	Revise
I85	0.6248	Easy	0.8707	0.3203	0.5504	Very Good	0.518	0.817	Retain
I86	0.4049	Moderately Difficult	0.3975	0.4141	–0.0166	Negative	–0.058	0.827	Reject
I87	0.274	Difficult	0.3407	0.1914	0.1493	Poor	0.146	0.823	Revise
I88	0.5236	Moderately Difficult	0.6814	0.3281	0.3533	Good	0.336	0.82	Retain
I89	0.2391	Difficult	0.2839	0.1836	0.1003	Poor	0.093	0.824	Reject
I90	0.438	Moderately Difficult	0.4984	0.3633	0.1351	Poor	0.128	0.824	Revise
I91	0.1466	Very Difficult	0.1451	0.1484	–0.0033	Negative	–0.049	0.826	Reject
I92	0.1832	Very Difficult	0.1956	0.168	0.0276	Poor	0.016	0.825	Reject
I93	0.1937	Very Difficult	0.2177	0.1641	0.0536	Poor	0.005	0.825	Reject
I94	0.3176	Difficult	0.388	0.2305	0.1575	Poor	0.156	0.823	Revise

Item	Difficulty	Difficulty Level	Upper Mean	Lower Mean	D	Discrimination	CITC	Alpha if Deleted	Decision
I95	0.3211	Difficult	0.429	0.1875	0.2415	Marginal	0.209	0.822	Revise
I96	0.2112	Difficult	0.2366	0.1797	0.0569	Poor	0.027	0.825	Reject
I97	0.3194	Difficult	0.3312	0.3047	0.0265	Poor	–0.029	0.826	Reject
I98	0.3944	Difficult	0.4259	0.3555	0.0704	Poor	0.017	0.825	Reject
I99	0.2182	Difficult	0.2303	0.2031	0.0272	Poor	–0.023	0.826	Reject
I100	0.2042	Difficult	0.2271	0.1758	0.0513	Poor	0.043	0.825	Reject

Note. Decision criteria used in this appendix were: Retain = difficulty from 0.20 to 0.75, discrimination ≥ 0.30, and corrected item–total correlation ≥ 0.15. Reject = negative discrimination, corrected item–total correlation < 0.10, or extreme difficulty (< 0.10 or > 0.90). All remaining items were classified as Revise.