p-ISSN 2085-7411
e-ISSN 2623-0380

Psychometric Evaluation of a Cognitive Test for Accreditation Assessors in Academic Libraries

Trimo Septiono[1,2], Awaludin Tjalla[1]
[1]Master's Program in Educational Research and Evaluation, Jakarta State University
[2]Directorate of Standardization and Accreditation, National Library of the Republic of Indonesia, Indonesia
*trimo.contact@gmail.com

Abstract

Limited empirical evidence exists regarding the validity of instruments used to assess evaluative competence in academic libraries accreditation. The 2025 revision of the Indonesian higher-education library accreditation instrument assigns the highest weighting to service components, requiring assessors to interpret relationships across indicators. This study examined the psychometric properties of a cognitive test instrument designed to assess such competence. A total of 32 items were administered to 30 assessors, and item characteristics were analyzed using classical test theory. Results indicate highly homogeneous response patterns: 90% of respondents scored in the high category, 81.25% of items were classified as easy, 46.88% had low discrimination indices, and 65.63% contained non-functional distractors. Although the KR-20 reliability coefficient was high (0.84), internal consistency was driven by response homogeneity rather than true variance in ability. These findings suggest that high internal consistency in cognitive assessment instruments may arise from response uniformity rather than reflecting actual variation in underlying ability, particularly in professional evaluation contexts.

Keywords: Library accreditation; Assessors; Higher-education Library accreditation instrument; Cognitive instrument; Item analysis.

Received: March 15, 2026. Revised: April 08, 2026
Accepted: April 21, 2026

Article Identity:
Septiono, T., & Tjalla, A. (2026). Psychometric Evaluation of a Cognitive Test for Accreditation Assessors in Academic Libraries. *Jurnal Ilmu Pendidikan (JIP) STKIP Kusuma Negara*, 18(1), 10-24.

INTRODUCTION

The transformation of academic libraries in the last decade has shifted the orientation of services from their traditional function as collection providers to strategic nodes in the institutional research ecosystem. The implementation of Research Data Management (RDM)-based services illustrates this shift through efforts to maintain the integrity of research data and the efficiency of research processes (Fadhli, 2018; May, 2019). Simultaneously, libraries have begun to develop scientific consultation services, data literacy, open access publication assistance, and repository management, therefore, the contribution of libraries is no longer passive and now directly involved in the production of academic knowledge.

However, this expanded role is not fully reflected in the evaluation instruments used during library accreditation. Analysis of accreditation data demonstrates that

most indicators are still oriented towards administrative compliance, while the dimension of research service performance has not been adequately accommodated (Haryono & Cahyono, 2020). This contrasts with findings regarding the contribution of institutional repositories to increasing academic publications and strengthening research culture in higher education (Nugroho, 2022). This misalignment indicates a gap between library service practices and the evaluation system used to assess their success. However, there is still limited empirical evidence on whether the instruments used to assess assessors' evaluative competence are able to measure such complexity accurately.

In this context, the position of academic libraries assessors becomes very strategic. Assessors not only examine documents or administrative requirements, but also interpret how institutions operationalize research services into library performance evidence. Accordingly, the quality of the assessment depends heavily on the assessor's ability to capture the dynamics of research services and translate them into an accreditation assessment. This makes measuring the evaluative abilities of assessors a fundamental requirement so that accreditation assessments do not stop at compliance aspects but are able to capture the role of libraries in enhancing the research capacity of institutions.

Preliminary analysis of accreditation data shows notable gaps between service performance and outcome indicators, particularly in innovation, reading engagement, and community literacy measures. In many cases, higher service scores are not followed by proportional outcomes. This suggests that the current evaluation framework does not fully capture the relationship between service performance and its intended impact.

Hence, the accreditation score based on the 2022 instrument seems to separate the assessment of services from the impact generated, therefore the interconnection between indicators is not apparent in the final assessment. The update of the 2025 accreditation instrument can be understood on this basis of these findings. When the weight of the service component is increased to the highest, namely 35%, service assessment will significantly influence the accreditation results.

However, this change in weighting requires assessors to not only examine the completeness or fulfilment of service elements, but also to determine whether service achievements are meaningfully related to innovation, user engagement in reading, and community literacy outcomes. This shift places greater emphasis on the assessor's ability to interpret relationships across indicators so that accreditation recommendations reflect both service performance and its broader impact on academic and literacy development.

This increased evaluative demand highlights a critical issue. Without an appropriate measurement instrument, the interpretation of inter-indicator relationships may rely on subjective judgment, which can lead to inconsistencies and reduced validity in accreditation outcomes. At the same time, empirical instruments specifically designed to assess this type of evaluative competence remain limited, indicating a gap between the complexity of assessment demands and the availability of suitable measurement tools. Based on these considerations, this study aims to evaluate the psychometric properties of a cognitive test designed to measure the evaluative competence of academic library accreditation assessors.

RESEARCH METHOD

This study adopted an initial psychometric validation approach, with a focus on item-level analysis using Classical Test Theory (CTT), to examine the quality of a cognitive assessment instrument for academic library accreditation assessors. The focus of the study is directed at analysing item characteristics and response patterns to obtain an empirical picture of the initial performance of the instrument before it is used in more extensive testing. A total of 30 assessors were involved as respondents through purposive sampling based on their involvement in the national academic libraries accreditation process. This number of respondents was considered adequate for the initial descriptive analysis of item quality and response trends, particularly in identifying potential psychometric weaknesses of the instrument.

Instrument quality analysis was conducted using a classical test theory approach, which included difficulty level, discriminating power, distractor effectiveness, item-total empirical validity, and internal reliability estimation using KR-20. The instrument consisted of multiple-choice items with five answer options for each question. Data analysis was performed using Microsoft Excel and Jamovi to calculate item statistics and reliability coefficients. The decision limit for empirical item validity was set at $r \geq 0.361$, based on the critical value at $\alpha = 0.05$ for $n = 30$ respondents. With this sample size, the findings are considered preliminary and mainly reflect initial item performance. The KR-20 reliability coefficient was interpreted using commonly accepted reliability interpretation criteria for internal consistency. This analysis was aimed at examining response variability and inter-item correlation as a basis for assessing the consistency of the instrument structure and identifying potential response bias that could affect score interpretation.

As stated by Creswell (2018), the descriptive quantitative approach in instrument development allows researchers to empirically examine item performance in a real measurement context, so that the results of the analysis can be used as a basis for refining construct representation and scoring mechanisms. In line with this view, this study not only describes statistical adequacy but also maps empirical risks, particularly response homogeneity and inter-item correlation, which could potentially influence the interpretation of assessors' cognitive abilities when the updated accreditation instrument is implemented in 2025.

RESULTS AND DISCUSSION

To get a clearer picture of how service performance relates to outcome indicators, Table 1 summarizes the gap patterns between K3 (service) and K7–K9 components in the 2022 library accreditation instrument. The table highlights that service achievements are not consistently reflected in innovation, reading engagement, and community literacy outcomes across different accreditation levels.

Table 1. Percentage of Gap Zones between K3 and K7–K9 (2022 Library Accreditation Instrument)

Type of Gap (K3 – Service)	A-Rated Libraries	B-Rated Libraries	Interpretive Notes (Concise Academic Form)
K3 → K7 (Innovation)	Extreme: 13.7% Moderate: 67.6% Negative: 18.7%	Extreme: 50.8% Moderate: 24.6% Negative: 24.6%	The higher proportion of extreme gaps in B - rated libraries suggests weak alignment between service performance and innovation outcomes. A - rated libraries indicate more stable linkage, reflected in the dominance of moderate gaps.
K3 → K8 (Reading Engagement)	Extreme: 25.2% Moderate: 63.3% Negative: 11.5%	Extreme: 63.1% Moderate: 15.4% Negative: 21.5%	Extreme gaps dominating B - rated libraries indicate that service performance does not correspond proportionally to users' reading engagement. A - rated libraries show a more proportional distribution across gap zones.
K3 → K9 (IPLM)	Extreme: 43.2% Moderate: 42.4% Negative: 14.4%	Extreme: 35.4% Moderate: 20.0% Negative: 44.6%	The dominance of negative gaps in B - rated libraries indicates that IPLM scores are not consistent with service performance. A - rated libraries show a more controlled and proportional distribution.

Zone Category	Definition
Extreme Zone	gap ≥ 10.00
Moderate Zone	9.99 ≥ gap ≥ 0
Negative Zone	gap ≤ 0

Source: Processed from Library Accreditation Big Data (2025).

The data on the gap between K3 and K7-K9 reflects that library service achievements are not directly correlated to innovation achievements, user reading habits, or IPLM. This lack of synchronization is clearly evident in libraries with a B rating, which show a very high percentage of extreme zones in K7 (50.8%) and K8 (63.1%) and a predominance of negative zones in K9 (44.6%). This pattern indicates that service values are not moving in the direction of the expected increase in results associated with them.

Overall, these patterns show that service performance does not always move in line with the expected outcomes, especially in libraries with lower accreditation levels. This raises an important question about how assessors interpret these relationships during the evaluation process. To better understand this issue, it is necessary to look at the assessment instrument itself, particularly how well it can capture differences in evaluative ability. The next section presents the initial instrument profile, focusing on item performance in the limited trial phase.

Initial Instrument Profile: Item Performance Profile in the Limited Trial Phase

The cognitive test instrument consists of 32 items with a maximum score of 32 and was administered to 30 university library assessors in the limited trial phase. The total scores ranged from 14 to 32, indicating a tendency toward higher score concentration, as presented in Table 2.

Table 2. Classification of Total Respondent Scores

Score Category	Score Range	Number Of Respondents	Percentage
High	26-32	27	90,0%
Medium	20-25	1	3,3%
Low	14-19	2	6,7%
Total		30	100%

Source: Author's data processing, 2025.

The score distribution in Table 2 illustrates that 90% of respondents are in the high category, while the medium and low categories cumulatively cover only 10% of respondents. This distribution pattern shows low score variance, which is consistent with a high difficulty index value ($p \geq 0.90$), even reaching $p = 1.00$ on some of the initial items. Thus, empirical data show that the instrument items tend to be too easy for the majority of respondents, so that the responses given are relatively uniform.

The homogeneity of the total scores has direct consequences for the estimation of psychometric parameters in the Classical Test Theory (CTT) framework. When the variance of the total scores is low and most of the responses are concentrated on correct answers, the covariance between the item scores and the total scores will decrease, thus limiting the instrument's ability to distinguish between groups of respondents with different ability levels. Under such conditions, a decrease in the strength of the relationship between item responses can occur even though the substance of the material is relevant. In other words, the limitations of the response pattern not only reflect the editorial quality of the questions, but also illustrate the limitations of the instrument in creating cognitive stimuli capable of eliciting differences in performance among respondents.

These empirical findings are consistent with previous studies. Rezigalla (2022) asserts that the dominance of easy items hinders the test's ability to detect variations in participants' abilities, while Wind (2022) shows that homogeneous score distributions in small samples reduce the sensitivity of item characteristic detection. Chhetri and Khanal (2024) also report that tests with a limited number of respondents often produce less varied score distributions even though the instruments have been systematically designed. The findings of Wahab et al. (2023)

and Moto et al. (2022) show a similar pattern, namely that early-stage instruments tend to show a dominance of easy items and weak discriminating power.

Thus, the distribution of total scores in this limited trial phase shows that the cognitive test instrument is not yet fully capable of revealing the variation in the cognitive abilities of academic libraries accreditation assessors. This condition has a direct impact on the research objective, which is to produce a test instrument that can provide sensitive and objective evaluative evidence of the variation in assessors' abilities in applying accreditation standards. Therefore, these empirical findings form the basis for refining the difficulty and complexity of the item stimuli before the instrument is retested on a more heterogeneous sample.

Item Response Variability through Difficulty Index Imbalance

The distribution of difficulty levels in the instrument shows a very uneven pattern. Of the total 32 items, empirical data shows that 26 items (81.25%) have a difficulty index ≥ 0.81, while 6 items (18.75%) are in the range of 0.61-0.80, with no items categorized as difficult (≤ 0.60). A p-index value of 1.00 on several items indicates that all assessors who were respondents answered the item correctly. This condition shows that the cognitive stimulus offered by the item was not challenging enough to elicit variations in ability among assessors.

Table 3. Distribution of Item Difficulty Index for PT Assessor Cognitive Test Instruments

Level of Difficulty	P Indeks Range	Number of Items	Percentage
Easy	≥ 0,81 – 1,00	26	81,25%
Medium	0,61 – 0,80	6	18,75%
Difficult	≤ 0,60	0	0%
Total		32	100%

Source: Author's data processing, 2025.

This disparity in difficulty levels is directly related to the findings in the previous subsection, namely that assessors' total scores tend to cluster in the high range. This means that the homogeneity of scores at the instrument level described in Table 3 reflects the homogeneity of responses at the item level. In other words, when most items are classified as easy, the instrument loses its diagnostic ability to reveal differences in assessors' cognitive abilities, so that the high scores obtained are not solely due to high ability but also due to the non-selective nature of the items.

From the perspective of educational evaluation, this condition risks producing biased measurement conclusions. Cognitive assessment instruments for assessors should function as selection tools capable of distinguishing assessors' levels of understanding of accreditation standards. These findings are consistent with the mapping of abilities presented by Yudiarso et al. (2025), which states that a proportional distribution of difficulty levels is necessary for responses to accurately reflect respondents' latent abilities.

Lintangesukmanjaya et al. (2025) also showed that the imbalance of item difficulty in the instrument can hinder the measurement model's ability to identify groups of respondents with different abilities. In the context of this instrument, the absence of difficult items indicates that the assessment has not touched on the

evaluative and diagnostic cognitive domains that should be at the core of the abilities of library accreditation assessors.

This finding is in line with Prasetiyo et al. (2022), that homogeneous high scores cannot be directly interpreted as evidence of high ability, but may be a consequence of items that are too easy. Consequently, an instrument dominated by easy items can produce misleading competency validation for accreditation agencies, as it does not consider the depth of analytical ability that is essential in the process of assessing the quality of academic libraries.

Thus, the imbalance in difficulty levels is not merely a statistical finding, but an indicator that this test instrument is not yet capable of facilitating measurements that are sensitive to variations in assessor abilities. Empirical data shows the need to reconstruct items by increasing the complexity of stimuli, using real accreditation case contexts, and adding items that measure reasoning and case diagnosis abilities to support the evaluative function of the instrument in the next stage of development.

Instrument Sensitivity: Item Discrimination Power and Assessor Ability Variation

The results indicate that 17 out of 32 items (53.13%) show acceptable discrimination, while 15 items (46.88%) have low discrimination values (≤ 0.20). This nearly balanced proportion indicates that the instrument does not yet consistently reveal differences in assessors' abilities to understand accreditation standards. This data is important because discriminating power is an indicator of the extent to which an instrument can select respondents' abilities, not merely count the number of correct answers.

Tabel 4. Distribution of Item Discrimination Power of PT Assessor Cognitive Test Instruments

Category	Range	Number of Items	Percentage
Adequate	≥ 0,21	17	53,13%
Inadequate	≤ 0,20	15	46,88%
Total of Items		32	100%

Source: Author's data processing, 2025.

These findings are closely related to the distribution of difficulty levels in the previous section, namely that 81.25% of items were in the easy category. When the cognitive stimulus is relatively simple, assessors' responses tend to be homogeneous, so that the items lose their ability to distinguish between high- and low-ability assessors. Empirical data show that most items in the non-discriminating category have a difficulty index close to the maximum value. Thus, the low discriminatory power is not solely due to weaknesses in the wording of the questions, but is a direct consequence of cognitively unchallenging stimuli.

This pattern appears consistent with previous research findings. In the case of the instrument reviewed by Wahab et al. (2023), items with a correct response rate of nearly 100% were unable to detect variations in participant ability, even though the total scores appeared high. A similar situation arises in this instrument, where high scores on some assessors are misleading because the items do not provide opportunities for ability differentiation. Fernanda and Hidayah (2020) also show

that the initial reliability of an instrument does not guarantee its discriminating power, especially when responses are not varied. This condition is reflected in this instrument, where stable scores cannot be assured to reflect the evaluative abilities of assessors.

A similar case was found by Moto et al. (2022), which showed that a significant proportion of no discriminative items can lead to erroneous conclusions about ability evaluation. This is relevant to the findings of this academic libraries accreditation assessment instrument, because high scores may originate from item characteristics rather than the ability to apply accreditation standards. Furthermore, Wind (2022) emphasized that in small-sample trials, low discriminative power increases the risk of incorrect ability inferences. These findings reinforce that the use of 30 assessors on an instrument with 46.88% no discriminative items can produce an impression of homogeneity of ability that is not factual.

Thus, the discrimination data at this stage not only indicates technical problems but also suggests that the instrument fails to facilitate the valid disclosure of assessors' evaluative abilities. High scores in the test cannot yet be positioned as evidence of assessors' competence in understanding and applying accreditation standards. The instrument requires reconstruction of stimuli on non-discriminative items so that the assessment can accurately reveal variations in ability in the next stage of development.

Effectiveness of Distractors and Accuracy of Assessor Inference

A recapitulation of the distractor power shows that the academic libraries assessor's cognitive test instrument is not yet fully capable of stimulating conceptual errors. Of the 32 items, 9 items (28.13%) had distractors that were not selected at all, and 12 items (37.50%) had distractors that were selected <5%. This means that nearly two-thirds of the items (65.63%) lost their diagnostic function because the incorrect options were not perceived as credible alternatives by the assessors. This condition indicates that the estimator of assessor ability based on the correct scores on these items is weak, because there are no opportunities for errors that can be used as indicators of misconceptions or immaturity in evaluative literacy.

Table 5. Recap of the Effectiveness of Distractors in Cognitive Test Instruments

Distractor Effectiveness Category	Operational Definition	Number of Items	Percentage
Does not function	There is a distractor with 0% of voters	9	28,13%
Not functioning effectively	There is distractor with <5% of voters	12	37,50%
Functioning effectively	All of distractors were selected ≥5%	11	34,38%
Total		32	100%

Source: Author's data processing, 2025.

This phenomenon is consistent with empirical patterns found in studies of distractors in various assessment contexts. Ansari et al. (2022) show that the presence of non-functional distractors indicates a flaw in the construction of options and threatens the credibility of scores because participants tend to immediately identify the correct answer without elaborating cognitively on the alternatives. This

mechanism is clearly reflected in items where distractors were not selected: the incorrect options appear too easy to eliminate, indicating that the distractors do not function as plausible alternatives. This pattern suggests that the instrument may not yet fully capture variations in evaluative reasoning among assessors.

Some of the academic libraries assessment data also shows that the distribution of incorrect options tends to be concentrated on only one or two distractors. This is consistent with the findings of Shakurnia et al. (2022) that an increase in the number of non-functional distractors correlates with an increase in item ease and a weakening of the item's ability to distinguish between test takers' abilities. In other words, the high number of easy items found in academic libraries assessment instruments does not solely reflect the high ability of the assessors, but also the weak competition among distractors. This situation supports the interpretation that inferences about ability cannot be based solely on correct scores, but must also consider the structure of incorrect responses that fail to appear due to the implausible design of distractors.

Interestingly, the data also found that items with a higher proportion of functional distractors tended to appear in items with a moderate level of difficulty. This pattern intersects with the report by Dayadaya & Sermona (2024) that items with more functional distractors tend to be at a higher cognitive level and encourage reasoning, rather than mere recall. Thus, the distribution of distractors in the academic libraries instrument can be interpreted as an indicator of item cognitive quality: functional distractors are more likely to appear when the item content requires evaluation of accreditation standards, rather than just recognition of procedural terminology.

Meanwhile, Syahrir et al. (2024) emphasize that the effectiveness of distractors is directly related to the instrument's ability to differentiate respondents based on their level of understanding rather than their capacity to eliminate irrelevant options. If distractors are too superficial or implausible, correct answers cannot be interpreted as representing substantive understanding, but rather as success in avoiding clearly incorrect options. In this context, the academic libraries assessment instrument appears to exhibit an identical pattern: the limited variation in distractor selection indicates that incorrect options may not sufficiently differentiate response patterns, so that correct responses may reflect recognition of procedural aspects rather than deeper evaluative reasoning. While these findings indicate weaknesses in distractor functioning, the interpretation of their impact on evaluative reasoning should be approached cautiously, as distractor analysis primarily reflects response patterns rather than direct evidence of cognitive processes.

Thus, the findings on the distribution of distractors in the assessor instrument are not merely statistical results, but an indication that this assessment has not stimulated the depth of reasoning required for the professional tasks of accreditation assessors. The acculturation of the findings in the above journals shows that weak distractors in the context of educational assessment consistently result in biased interpretations of ability. Therefore, improving distractors is crucial not as a cosmetic effort in question construction, but as an effort to maintain the validity of inferences and strengthen the credibility of the cognitive assessment of college library assessors.

Empirical Validity Structure: Consistency of Response Patterns Across Items
Of the 32 items analyzed, 22 items (68.75%) exceeded the item–total correlation threshold of 0.361. Nevertheless, reliance on threshold criteria alone is limited, and the magnitude of correlations should be considered alongside response patterns in interpreting item quality. The stable inter-item correlations observed in these items should be interpreted cautiously and cannot be assumed to fully represent construct coherence because the assessor response profiles in the score table show a pattern of correct answers concentrated in the high score range. This homogeneity of responses, which was also apparent in the difficulty index and distractor effectiveness, can produce high correlations mathematically even though the observed response patterns may not necessarily reflect variation in evaluative processes between items.

Thus, the empirical validity reflected in the correlation coefficient may also be influenced by response homogeneity, indicating the need for cautious interpretation of whether the observed consistency reflects underlying ability or response patterns. This interpretation is limited to observed response patterns and does not constitute a direct test of construct-level validity.

Table 6. Empirical Validity of PT Assessor Cognitive Test Instruments

Item Validity Category	Decision Limit r-count	Number of Items	Percentage	Psychometric Implications
Empirical valid	r-count ≥ 0.361	22	68,75%	high consistency of responses between items, but it is necessary to analyze whether the consistency stems from the construct or the homogeneity of responses
Empirical unvalid	r-count < 0,361	10	31,25%	the pattern of responses between items is not consistent, potentially measuring different constructs/ambiguity of stimuli
Total		32	100%	

Source: Author's data processing, 2025.

Conversely, 10 items (31.25%) had r-counts below the significance threshold. Responses to these items showed variations that were not consistent with the response patterns on other items. These inconsistent variations in response suggest that these items did not elicit the same cognitive structure as that triggered by the majority of valid items. Judging from the previous discrimination and distractor data, this variation in responses did not always arise because of the assessors' diverse abilities, but could also be because the item stimuli did not accurately map the evaluative competencies that were the construct of the instrument.

The case of correlations that appear valid in most items and low correlations in these 10 items can be interpreted in line with the thinking in the development of item-total correlation-based instruments -total, that consistency between items can

be formed not because of construct alignment, but because of homogeneous responses due to easy stimuli or weak distractors, as explained in the study of reliability and response consistency analysis (Chhetri & Khanal, 2024). The correlation pattern in this instrument shows a similar mechanism: consistency arises because there is no sufficient variation in responses to "shake" the item-total correlation.

Similarly, the principle of item-total correlation interpretation in item analysis shows that when an item produces a high correlation but is not accompanied by differentiation of responses among respondents, the correlation may indicate a technical relationship between items, rather than a true latent construct relationship, as explained in the framework for interpreting item-total correlations (Rezigalla, 2020). This mechanism provides a perspective that the consistency of responses on the academic libraries assessment instrument can be understood as a reflection of the structure of the stimulus, not ability.

Furthermore, the variability of correlations between items in this instrument can be understood as a consequence of the uneven distribution of assessor abilities and item stimulus characteristics, especially when some items are unable to capture differences in evaluative concept capacity between respondents—a mechanism described in the response structure and item suitability approach in a professional context (Lintangesukmanjaya et al., 2025). Thus, the empirical validity of this instrument reflects the mathematical coherence between items that is not yet fully rooted in construct structural alignment. The high correlation of many items can be better understood as the result of response homogeneity due to overly easy stimuli and nonfunctional distractors, while low correlations indicate the failure of items to elicit the same evaluative structure. The implication is that the interpretation of total scores does not guarantee that high scores reflect the evaluative abilities of assessors; rather, they reflect a stimulus profile that is unable to demonstrate the diversity of existing abilities.

Internal Reliability of Instruments Based on KR-20: Stability of Assessor Ability Inferences

The KR-20 coefficient of 0.8446 was formed from a total score variance of 13.61 and $\Sigma(pq)$ of 2.64, indicating that the covariance between items moved uniformly. When linked to the findings structure in the previous chapter, it can be seen that the mathematical components that form reliability move in line with a homogeneous response pattern, namely that the majority of items are in the easy difficulty category, have low discriminating power, and distractors are not selected. These three conditions narrow the score distribution because most assessors choose the correct answer in a relatively similar way, resulting in low inter-assessor variance. In this type of score structure, a high reliability coefficient can be established due to the consistency of response direction, not because the instrumentation is able to express latent ability variation.

Table 7. KR-20 Reliability Results of PT Assessor Cognitive Test Instruments

Calculation Components	Results
Number of respondents	30
Number of valid items	22
Total score variance	13,61
Σ(pq)	2,64
KR-20 value	0,8446
Reliability category	Significantly high

Source: Author's data processing, 2025.

The high internal consistency observed at this trial stage should be interpreted cautiously. While it may be associated with the homogeneous response patterns observed in the data, reliability coefficients do not in themselves explain the source of score consistency, as noted in previous studies (Wind, 2022). The principle that item covariance can increase when item stimuli guide responses in one direction has been discussed in the classical item analysis framework by Miller et al. (2012). Meanwhile, Wind (2022) shows that in small samples, homogeneous responses can mask item weaknesses, because item covariance remains high even though the stimuli do not clearly map abilities.

When considered alongside the observed response patterns, the KR-20 value of 0.8446 indicates a high level of internal consistency; however, it does not in itself distinguish whether this consistency arises from similarities in ability or from uniform response patterns influenced by item characteristics. This suggests that the observed consistency in total scores may be associated with the regularity of response choices across items, as seen from the small Σ(pq), than a characteristic of regularity of ability. The resulting inter-item correlations may reflect similar response patterns across items, rather than necessarily indicating consistent expression of underlying ability.

Thus, when the KR-20 value of 0.8446 is placed in the context of the previous findings, the coefficient shows that internal consistency in this instrument does not serve as evidence of construct reliability, but rather as an indicator that responses move in the same channel due to stimuli that do not elicit different responses. The correlation between items is formed not because evaluative abilities are stably distributed, but because the majority of items encourage homogeneous responses, as indicated by easy difficulty, small Σ(pq), and distractors that do not function. Thus, high reliability in the context of this data indicates regularity of responses, not regularity of abilities; the stability of the coefficient is more a reflection of the structure of the items than a reflection of the stability of the assessors' competencies.

Structural Consistency Based on Psychometric Index Convergence

Consistency of Construct Structure Based on Empirical Data Convergence The consistency of assessor response patterns can be seen from the convergence of statistical indicators that reinforce the direction of homogeneous responses. Of the 32 items analyzed, 26 items (81.25%) were classified as easy and 21 items (65.63%) had distractors that were not selected. These two indicators reduce the possibility of incorrect response variations. The statistical effect is seen in the low Σ(pq) (2.64) and narrow total score variance (13.61), indicating limited opportunities for differentiation of responses between assessors.

This limitation in response variation continues in the discriminating power. A total of 15 items (46.88%) showed low discriminating power, which means that assessors with high and low total scores tend to choose the same answers. When correct responses were dominant in all score groups, item-total correlations appeared valid for 22 items (68.75%), but these correlations reflected the repetitiveness of response directions rather than substantive relationships between items. The stable inter-item covariance pattern then produced a KR-20 reliability coefficient of 0.8446.

These indicators collectively form a response mechanism that focuses on the stimulus item rather than on variations in the evaluative abilities of assessors. Item structures that are too easy and passive distractors are sources of response homogeneity. In distractor analysis studies, this situation is described as a condition where incorrect choices do not play a diagnostic role, so responses are directed toward one correct option (Shakurnia et al., 2022; Dayadaya & Sermona, 2024). This homogeneity of responses, which narrows the variance in scores, can result in high reliability at the stage when the stimulus does not separate latent abilities (Chhetri & Khanal, 2024).

This tendency is reinforced by the relationship between high difficulty, low discriminating power, and passive distractors as shown in item quality analysis research (Shakuri et al., 2019; Sumagang et al., 2017). When these three elements are present simultaneously, the covariance between items increases not because of the functional relationship between ability indicators, but because the stimulus structure guides responses in a uniform direction. This condition is evident in the data from the academic libraries assessor's cognitive assessment instrument. Thus, the observed consistency in the instrument structure appears to be associated with the regularity of response patterns influenced by item characteristics, rather than necessarily reflecting stable differences in evaluative ability. All empirical indicators, including the proportion of easy items, non-functional distractors, low discriminating power, small $\Sigma(pq)$, narrow score variance, valid item-total correlation, and high KR-20, consistently point to a pattern of homogeneous responses. This pattern may be interpreted as structural in nature, suggesting that score consistency is influenced by stimulus characteristics, rather than providing direct evidence of evaluative ability.

CONCLUSION

The results of this study indicate that cognitive assessment instruments for academic libraries accreditation assessors are not yet fully capable of adequately representing variations in evaluative abilities. The tendency toward homogeneous response patterns, the dominance of items with high difficulty levels (easy category), limited discriminating power, and low distractor effectiveness indicate that the stimulus structure of the instrument encourages response uniformity rather than facilitating ability differentiation. This condition has a direct impact on the formation of statistically high score consistency, but should be interpreted cautiously and not assumed to directly represent the reliability of assessor ability constructs.

This study shows that high internal consistency in a cognitive assessment instrument does not always reflect real differences in evaluative ability, but may

also be influenced by uniform response patterns. It also highlights the need to interpret reliability together with item characteristics, such as difficulty, discrimination, and distractor functioning, so that score-based conclusions can be made more carefully.

Building on these findings, it is important to balance item difficulty, discriminating power, and distractor quality in developing cognitive assessment instruments, especially in professional contexts that require evaluative reasoning. Therefore, instrument refinement should focus on increasing the complexity of stimuli based on real accreditation contexts, designing distractors that represent common misconceptions among assessors, and testing instruments on more diverse respondent groups. These steps are necessary for the instrument to function as a valid and sensitive measuring tool in supporting the implementation of academic libraries accreditation in line with the application of the updated accreditation instrument in 2025.

REFERENCES

Ansari, M., Sadaf, R., Akbar, A., Rehman, S., Chaudhry, Z. R., & Shakir, S. (2022). Assessment of distractor efficiency of MCQs in item analysis. *Professional Medical Journal, 29*(5), 730–734. https://doi.org/10.29309/TPMJ/2022.29.05.6955

Chhetri, D. B., & Khanal, B. (2024). A pilot study approach to assessing the reliability and validity of relevancy and efficacy survey scale. *Janabhawana Research Journal, 3*(1), 35–49. https://doi.org/10.3126/jrj.v3i1.68384

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Dayadaya, R. M. N., & Sermona. (2024). Analysis of distractor efficiency of the multiple-choice comprehensive examination for technical-vocational pre-service students. *International Journal for Multidisciplinary Research, 6*(6), 1–6. https://doi.org/10.36948/ijfmr.2024.v06i06.32251

Fadhli, R. (2018). Peran perpustakaan perguruan tinggi dalam research data management untuk mendukung scholarly communication. *Khizanah Al-Hikmah: Jurnal Ilmu Perpustakaan, Informasi, dan Kearsipan, 6*(2), 122–131. https://doi.org/10.24252/kah.v6i2a4

Fernanda, J. W., & Hidayah, N. (2018). Analisis kualitas soal ujian statistika menggunakan classical test theory dan rasch model. *SQUARE: Journal of Mathematics and Mathematics Education, 2*(1), 49–60. http://dx.doi.org/10.21580/square.2020.2.1.5363

Haryono, B. S., & Cahyono, D. (2020). Implementasi kebijakan standar nasional perpustakaan perguruan tinggi di Perpustakaan Universitas Negeri Malang. *BACA: Jurnal Dokumentasi dan Informasi, 42*(2), 179–191. https://doi.org/10.14203/j.baca.v41i2.640

Ikhsanudin, I., Novaliah, N., Hidayatullah, H., & Almizi, M. (2023). A practical using of the QUEST program to analyze the characteristics of test items in educational measurement. *JISAE (Journal of Indonesian Student Assessment and Evaluation), 9*(1), eP2307. https://doi.org/10.21009/jisae.v9i1.31163

Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature, and Culture, 6*(2), 256–269. http://dx.doi.org/10.30659/e.6.2.256-269

Lintangesukmanjaya, R. T., Rizka, A. M., Damarsha, A. B., & Dwikoranto. (2025). Rasch model analysis: Case study in SDGs trend point 7 in physics learning based domicile. *The Journal of Indonesia Sustainable Development Planning, 6*(2), 173–184. https://doi.org/10.46456/jisdep.v6i2.608

May, L. (2019). Dynamic research support for academic libraries [Book review]. *Journal of Librarianship and Scholarly Communication, 7*(1), eP2307. https://doi.org/10.7710/2162-3309.2307

Moto, A., Musyarofah, L., & Taufik, K. S. (2022). Developing item analysis of teacher-made test for summative assessment of seventh grade of SMPN 8 Komodo in academic year 2020/2021. *Budapest International Research and Critics Institute Journal, 5*(1), 5305–5315. https://doi.org/10.33258/birci.v5i1.4237

Nugroho, P. A. (2022). Korelasi antara jumlah publikasi dosen dengan penggunaan layanan RemoteXs Perpustakaan Universitas Airlangga. *Daluang: Journal of Library and Information Science, 2*(1), 62–70. https://doi.org/10.21580/daluang.v2i1.2022.10060

Rezigalla, A. A. (2022). Item analysis: Concept and application. In M. S. Firstenberg & S. P. Stawicki (Eds.), *Medical education for the 21st century*. IntechOpen.

Setiawan, F., Andhika, B., & Yudiarso, A. (2025). Evaluation of validity and reliability of the eight-item Body Shape Questionnaire (BSQ-8C) Indonesian version: Mixture Rasch model approach. *Jurnal Sains Psikologi, 14*(1), 39–55. http://dx.doi.org/10.17977/um023v14i12025p39-55

Shakurnia, A., Ghafourian, M., Khodadadi, A., Ghadiri, A., Amari, A., & Shariffat, M. (2022). Evaluating functional and non-functional distractors and their relationship with difficulty and discrimination indices in four-option multiple-choice questions. *Education in Medicine Journal, 14*(4), 55–62. https://doi.org/10.21315/eimj2022.14.4.5

Wahab, A., Hasibuan, A., Siregar, R., Risnawaty, & Ningsih, T. Z. (2023). Item analysis of final examination questions for social studies in junior high schools through the ITEMAN program. *Journal of Educational Research and Evaluation, 7*(3), 526–536. https://doi.org/10.23887/jere.v7i3.59239

Wind, S. A. (2022). Identifying problematic item characteristics with small samples using Mokken scale analysis. *Educational and Psychological Measurement, 82*(4), 747–756. https://doi.org/10.1177/00131644211045347

Yudiarso, A., Ardhiani, I. W., Surya, R., Watimena, F. Y., & Kanzaki, M. (2025). Psychometric properties of the 18-item Indonesian mental toughness questionnaire using the Rasch model and machine learning. *Psikohumaniora: Jurnal Penelitian Psikologi, 10*(1), 747–756. https://doi.org/10.21580/pjpp.v10i1.23214