Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Development and validation of a college admission test for a higher education institution in the Cordillera Administrative Region, Philippines

Donato O. Abaya *

Department of Psychology, Ifugao State University, Lamut, Ifugao, Philippines

ARTICLE INFO

Article history:
Received 13 March 2025
Received in revised form
4 August 2025
Accepted 19 August 2025

Keywords:
College admission
Verbal reasoning
Numerical ability
Test validity
Test reliability

ABSTRACT

This study aimed to design a college admission test for a higher education institution in the Cordillera Administrative Region of the Philippines and to evaluate its psychometric properties. The initial version of the test included 120 items each for Verbal Reasoning and Numerical Ability, which were reviewed by experts and assessed for validity. After revisions, the final version contained 60 items for each area. The Item-Content Validity Index was 0.93, showing a high level of content validity. Reliability testing showed a coefficient of 0.80 for Verbal Reasoning and 0.65 for Numerical Ability, with an overall reliability of 0.76, indicating moderate but acceptable reliability. Exploratory Factor Analysis (EFA) confirmed that Verbal Reasoning had a three-factor structure, with all factor loadings above 0.40, supporting construct validity. Numerical Ability was found to represent a single factor, suggesting it measured one main ability. To check concurrent validity, the new test was given alongside the Otis-Lennon School Ability Test (OLSAT), and results showed strong positive correlations between similar subtests ($r = .71$, $p < .01$), supporting its criterion-related validity. Test norms were created using z-scores, IQ equivalents, and stanine scores. Overall, the findings show that the developed college admission test is a valid, reliable, and regionally appropriate tool for selecting incoming students.

1. Introduction

Testing has long been an essential tool in every educational institution. Through testing, educators could have a sound understanding of the students' aptitudes, interests, abilities, and needs, which could be useful in their decision-making in any of their academic undertakings. While non-standardized tests are common, standardized tests play a critical role in education because they are oftentimes utilized in diagnosing students' strengths and weaknesses, in enhancing students' motivation and personality, and most importantly, in planning the best ways to help students learn and adjust in their academic endeavors. Globally, tests like the Scholastic Aptitude Test (SAT) and American College Test (ACT) are used as university admission benchmarks to assess competencies (Maruyama et al., 2024). Similarly, in the Philippines, higher education institutions administer their entrance examinations to screen prospective students following the abolition of the National College Entrance Examination (Magno and Gonzales, 2011). However, concerns have been raised regarding the cultural and contextual relevance of foreign-made standardized tests in accurately assessing Filipino students' academic capabilities (Gatcho et al., 2024). In their book, Cerado and Garcia (2022), have pinpointed the relevance of tools for a specific locality or region in academic assessment. Hence, this highlights the importance of establishing a localized admission test specifically among higher educational institutions in the Cordillera Administrative Region of the Philippines.

Scholars have emphasized the importance of culturally relevant assessments to enhance test reliability and validity (Sternberg, 2018; Lee et al., 2011; Han et al., 2019). Likewise, Zhou et al. (2021), Green et al. (2025), and Dechavez (2024) found in their studies that using contextually appropriate tests yielded more precise results in evaluating students specifically on academic achievement tests.

In understanding why some standardized tests do not yield more precise results, studies such as those of Ozturgut (2011) and Penn (2023) argued that using non-localized standardized tests can

* Corresponding Author.
Email Address: xerdon17@gmail.com
https://doi.org/10.21833/ijaas.2025.09.020
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-7733-5031

present difficulties, including unsuitable norms among individuals who are taking the test and cultural biases. In the Philippines, Western standardized admission tests are widely used in some local universities; however, it does not fully encompass their sole purpose of effectively assessing the skills of Filipino test takers, specifically their academic performance (Magno and Gonzales, 2011). Considering the emerging issue when it comes to how precise and effective foreign-made standardized tests are, Pearce et al. (2015) underscored the necessity for universities and colleges in the Philippines to contextualize their assessment materials that are tailored to their academic requirements. This is to avoid the likelihood of inaccurately quantifying the abilities and educational experiences of Filipino students.

Hence, this research aimed to develop a college admission test with robust psychometric properties tailored to the required competencies of students in higher educational institutions in the Cordillera Administrative Region of the Philippines. The results of this study are expected to be relevant to the educational and psychological assessment fields by providing an empirically validated, reliable, and culturally sensitive admission test. This localized admission test may provide a more equitable and contextually appropriate measure of students' academic potential, addressing the limitations of foreign-made admission tests. Additionally, this study may provide substantial information to make data-driven decisions regarding student admissions among university officials. The result of the study may contribute to the growing body of knowledge on the relevance of locally developed college admission tests in fostering a more effective and inclusive higher education system in the Philippines.

2. Methodology

2.1. Respondents

The study utilized a descriptive method, which was conducted in eleven feeder senior high schools in the Cordillera Administrative Region, Philippines. The developed tests were initially piloted in large schools, involving 205 Grade 12 senior high school students. A second pilot test was subsequently conducted in smaller schools, with the participation of 169 Grade 12 senior high school students. While this sample size meets the general rule-of-thumb for psychometric validation studies, such that at least 5-10 respondents per item for item analysis and factor analysis (Hair et al., 2014), potential sampling bias is acknowledged. Most participants came from public schools, which may limit the representativeness of the population, particularly with respect to private school students from more urbanized regions. Additionally, although formal power analysis was not performed a priori, post-hoc evaluation suggests that the sample size was adequate to detect medium effect sizes (Cohen's d = 0.5) with 80% power in basic correlations and factor analysis (Faul et al., 2009). However, the absence of stratified or random sampling introduces potential selection bias, as schools were chosen based on administrative convenience and accessibility. This limitation may influence generalizability, and future research may incorporate multi-stage or stratified random sampling to enhance external validity and statistical power.

Upon the establishment of the item-content validity index and reliability coefficient of the developed college admission test, 1,700 incoming first-year students took the Otis-Lennon School Ability Test (OLSAT) and the developed college admission test simultaneously. This was conducted to establish the concurrent validity of the newly developed college admission test.

2.2. Procedures

The development and validation followed an 11-step process, including Exploratory Factor Analysis (EFA) and concurrent validity.

Using these steps, 1) the researcher reviewed a variety of books, journals, and internet sources regarding standardized admission tests and validation of tests to conceptualize what kind of admission test to develop; 2) the researcher created a table of specifications based on the objectives of the topic contents in basic and advanced English subjects for Verbal Reasoning and in basic and advanced Mathematics subjects for Numerical Ability. After this, item pooling on verbal reasoning and numerical ability from selected English and Mathematics teachers from a secondary school was done. Meanwhile, other test items were taken from textbooks, journals, manuals, and internet sources among others; 3) the constructed tests were reviewed and critiqued by the English and Mathematics experts from a University in the Cordillera Administrative Region as to the following components: a) Table of Specification; b) Content; and c) Organization; and d) Materials and Resources. Also, items were rated as to whether it is favorable or unfavorable using the Experts Assessment Form which became the basis for computing the Item-Content Validity Index (I-CVI) of the constructed tests; 4) the first pilot testing was conducted in big secondary schools; 5) the data from the first pilot testing were summarized and subject to item analysis to determine which items were accepted, modified or rejected; 6) after the item analysis, EFA was conducted to examine the underlying factor structure of the test based on empirical data from the first try-out. This helped validate whether the test items grouped according to the intended constructs; 7) the revised test items underwent second pilot testing in smaller schools; 8) and 9) the results of the second pilot testing became the bases for computing reliability coefficient of the tests; 10) a concurrent validity test was conducted to further determine how well the developed college admission test agrees with a well-established, validated measure particularly the OLSAT. The OLSAT and the newly developed and validated college admission test were administered simultaneously to the incoming first-year students; and 11) establishing norms based on the data gathered from the second pilot testing.

2.3. Data analysis

Item-content validation of the test items was conducted by a panel of experts using the Experts' Assessment Form, which followed a dichotomous rating system: Favorable (F+) for relevant items, assigned a score of +1.0, and Unfavorable (F) for non-relevant items, assigned a score of +0.0 (Zamanzadeh et al., 2015). There were five expert evaluators in this study, and it was set that the minimum level of agreement among five experts was more than 0.80, meaning at least four of them needed to concur for an item to be included in the final instrument and classified as valid. A computed I-CVI of 0.78 or higher is acknowledged by Shi et al. (2012) as an excellent content validity index. Similarly, Polit et al. (2007) reinforced that an I-CVI of 0.78 or higher is indicative of good content validity, specifically when there are three or more expert validators. Consequently, items with an I-CVI that is below 0.80 were discarded in this study.

In the process of analyzing the items, the researcher utilized the upper-lower 27% rule as an underpinning in determining the discrimination and the level of difficulty of each item. This involves comparing the top and bottom 27% of examinees. Such a procedure is supported by Rudolph et al. (2019), as it was explained that the method is contextually relevant to computing the discrimination index.

A reliability analysis was conducted using the Kuder–Richardson Formula 20 (KR-20) to assess internal consistency, which is appropriate for instruments with binary responses (e.g., right/wrong items). This method has been successfully applied in previous studies, such as the validation of a health literacy measurement tool for late school-aged children, where KR-20 coefficients exceeded the acceptable threshold of 0.70 (Park and Kim, 2021). A coefficient index of 0.80 or higher indicates high reliability and high acceptability, 0.50 to 0.79 moderate and acceptable, and less than 0.49 indicates low reliability and unacceptable. This was used in this study to determine and set the reliability coefficient of the developed college admission test.

To establish construct validity, an EFA was performed separately for each subtest-Verbal Reasoning and Numerical Ability. The analysis utilized principal axis factoring with varimax rotation to identify latent dimensions within each test component. Factor retention was based on the Kaiser (1974) criterion (eigenvalues > 1) and inspection of scree plots. All assumptions, including sampling adequacy and correlation matrix suitability (via KMO and Bartlett's Test), were first checked to ensure factorability. To assess concurrent validity as a form of criterion-related validity, respondents were asked to take two tests in the same testing session: the newly developed college admission test and the OLSAT, a widely recognized standardized test of school ability. Correlating scores from both instruments allows for determining whether the new test aligns with an established measure administered under identical conditions. This procedure follows best practices in psychological test validation, wherein test scores are compared to concurrent benchmark known to assess similar constructs (DeVellis, 2017; Messick, 1995; Tavakol and Wetzel, 2020). A Pearson correlation coefficient was computed for the total scores of the two tests.

Finally, norms were analyzed and set using z-scores, IQ scores, College Entrance Examination Board (CEEB), and Stanine, which are the standards for educational measurement-based norms.

3. Results and discussion

3.1. Initial content validity

Table 1 presents the Item-Content Validity Index (I-CVI) test results for the developed college admission test. Based on expert validation, the relevance and appropriateness of the test items are assessed. According to Kyriazos and Stalikas (2018), expert evaluators who are subject matter specialists perform a significant role in determining the content validity of the test.

Looking into the details, it can be gleaned from the result that the college admission test yielded an overall I-CVI score of 0.87. Specifically, the I-CVI values range from 0.85 to 0.99 across all five subtests, which is interpreted as favorable. The results confirm that item validity exceeds the 0.78 threshold for I-CVI when the panel consists of five or more experts (Polit and Beck, 2006; Lynn, 1986). Interestingly, Vocabulary (0.99) and Language Usage (0.97) achieved near-perfect agreement among experts. This reflects that the items under these components are highly aligned with the intended competencies assessed.

The slightly lower I-CVI values in the Analogy (0.85) subtest, while still favorable, suggest room for further refinement. This subtest may include items that, while relevant, may have been subject to varied interpretations by experts due to contextual or cognitive complexity. This aligns with the observations of Zamanzadeh et al. (2015), who emphasized that even when I-ICVI values are above the acceptable threshold, critical review of items with lower scores should still be undertaken to ensure clarity and alignment with construct definitions.

On the other hand, the high I-CVI for the Numerical Ability (0.93) is noteworthy, as it reflects a strong agreement among experts on the relevance of the items to assess quantitative reasoning and problem solving, a critical component of academic preparedness in STEM-related programs.

The overall findings highlight a high level of content suitability and relevance, which means a strong content validity. It affirms that conducting content validation is a fundamental basis in developing standardized psychological tests (Almanasreh et al., 2019; Yusoff, 2019). Such agreement among experts suggests that the developed items align well with regional and local academic standards and contextual expectations. The implication is that the tool is likely to yield valid insights into college readiness when used for screening. However, periodic expert review may be conducted to sustain this validity amid changing curricula and the educational landscape in higher education institutions.

Table 1: Summary of I-CVI of the college admission test

Tests	I-CVI	Interpretation
Verbal reasoning	0.87	Favorable
Analogy	0.85	Favorable
Language usage	0.97	Favorable
Vocabulary	0.99	Favorable
Numerical ability	0.93	Favorable
Over-All I-CVI	0.87	Favorable

3.2. First pilot testing for item analysis

Table 2 presents the summary of item analysis results for each subtest of the developed College Admission test. The analysis focused on identifying items that met the acceptable psychometric criteria – specifically item difficulty, item discrimination, and distractor efficiency – and subsequently determined which items were retained or discarded.

As shown in Table 2, after item analysis, 21 out of 40 items on Verbal Reasoning were retained, while 19 items were discarded for not meeting the intended measurement criteria, specifically the discrimination and difficulty level. In terms of Language Usage, 23 items out of 40 were retained, and 17 items were discarded while a total of 21 items were retained, and 19 items were removed under the Vocabulary section. Further analysis was conducted on the Numerical Ability section, and it was found that 60 out of 120 items were retained, while 60 items were discarded. The validity of the Numerical Ability subsection of the test in assessing mathematical competencies is enhanced by removing the 60 items.

The findings suggest that through systematic item analysis, the quality of the test could be improved, and it ensures that each retained item contributes to the reliability and validity of the subtests. The studies of Quaigrain and Arhin (2017), Yahia (2022), and Ashraf and Jaseem (2020) affirmed that conducting item analysis is pivotal in identifying and eliminating ambiguous items. The retention rates – ranging from approximately 52.5% to 57.5% for the verbal-related subtests- are within the expected range in large-scale test development, where initial item pools are purposely broad to allow for refinement (Haladyna and Rodriguez, 2013).

While reliability is enhanced by removing half of the Numerical Ability items, this significant discard rate of the items warrants further examination. The 50% rejection rate may suggest that a large proportion of the original items did not meet desired psychometric thresholds, possibly due to poor item discrimination or misalignment with the targeted constructs. According to Ferrando and Morales-Vives (2023), there are factors that contribute to this high discard rate, such as redundancy or biased distribution. However, these retained items may likely represent a refined and more reliable item set. This affirms the contention of Tavakol and Doody (2015) that refining item quality minimizes subjective errors, and it bolsters the reliability of psychological tests.

The results may impact the quality and integrity of the College Admission Test. The high number of discarded items across subtests underscores the importance of thorough pilot testing and psychometric evaluation prior to final deployment. It also indicates that future test development efforts may include multiple rounds of field testing and revisions to minimize item rejection and improve initial item pool quality.

Table 2: Summary of item analysis for each of the tests

Test	Total items retained	Total items discarded
Verbal reasoning	65	55
Analogy	21	19
Language usage	23	17
Vocabulary	21	19
Numerical ability	60	60

3.3. Exploratory factor analysis

Conducting the EFA yielded a valid and interpretable factor structure for both the Verbal Reasoning and Numerical Ability components of the test. Using Principal Axis Factoring with Varimax rotation, the analysis of the Verbal Reasoning component retained three factors consistent with the test's theoretical subdomains – Analogy, Vocabulary, and Language Usage. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was 0.84, indicating meritorious suitability for factor analysis (Kaiser, 1974), while Bartlett's Test of Sphericity was significant, $\chi^2(7140) = 5300.65$, $p < .001$, confirming that the correlation matrix was not an identity matrix and therefore factorable. The eigenvalues for the three retained factors were 13.42, 3.77, and 4.90, each exceeding the Kaiser (1974) criterion of 1.0.

Factor 1 (Analogy) accounted for 22.4% of the variance, Factor 2 (Language Usage) contributed 6.3%, summing to a total of 36.9% of the variance explained, and Factor 3 (Vocabulary) explained 8.2%. These three factors reflect distinct dimensions of verbal reasoning skills. Items that loaded into the Analogy factor ranged from 0.58 to 0.83, Language Usage from 0.50 to 0.79, and Vocabulary from 0.54 to 0.81. These findings are consistent with Nagy et al. (2012), who argued that academic language consists of multiple interconnected domains that must be separately evaluated to support language development and educational equity. The consistency between expert item evaluations (with I-CVI values of .85 for Analogy, .97 for Language

Usage, and .99 for Vocabulary) and the EFA-derived factor solution reinforces the test's content and construct validity. Triangulating expert judgement with empirical validation enhances test defensibility and increases interpretability for instructional use. Furthermore, the alignment between verbal subscales and empirical loadings affirms the diagnostic value of the Verbal Reasoning section, enabling detailed insights into students' strengths and areas for development.

Table 3 presents the EFA results for Verbal Reasoning and Numerical Ability.

Table 3: Exploratory factor analysis results for verbal reasoning and numerical ability

Test	Eigenvalue	Variance explained	Factor loading range
Verbal reasoning	22.09	36.9	.50-.83
Analogy	13.42	22.4	.58-.83
Language usage	3.77	6.3	.50-.79
Vocabulary	4.90	8.2	.54-.81
Numerical ability	11.85	19.75	.45-.88

In contrast, the EFA on the Numerical Ability section produced a one-dimensional factor solution, indicating that all items measured a single underlying construct of general quantitative reasoning. The KMO for Numerical Ability was 0.81, also considered meritorious, and Bartlett's Test was significant χ^2 (7140) = 4921.88, p < .001, supporting the suitability of the data for factor analysis. Only one factor had an eigenvalue greater than 1, specifically 11.85, and it explained 19.75% of the total variance. Factor loading for the retained 60 numerical items ranged from 0.45 to 0.88, supporting their strong alignment with the general quantitative reasoning construct. These findings, supported by Primi et al. (2010) and van der Maas et al. (2006), affirmed that mathematical performance across diverse items (e.g., problem-solving, number sense, and data interpretation) is driven by a unified dimension of fluid intelligence. The mutual reinforcement of mathematical skills contributes to the emergence of a general mathematical reasoning factor, a theory supported by the single-factor result found in this analysis. The alignment of this one-dimensional empirical result with the experts' I-CVI score of .93 for Numerical Ability underscores the coherence of the test design.

The implication of these findings is twofold. First, the three distinct factors in the Verbal Reasoning test allow for sub-score reporting, which provides richer diagnostic feedback for educational planning and remediation. This is particularly relevant for identifying areas where students may require intervention-whether in abstract verbal reasoning, vocabulary acquisition, or grammatical precision. Second, the one-dimensional structure of the Numerical Ability test justifies reporting a single composite score, simplifying score interpretation for admission purposes while ensuring high construct validity. As emphasized by DeVellis (2017), maintaining conceptual clarity and factorial integrity in educational assessments enhances both the interpretability and usability of test results for stakeholders. The general findings support the construct validity of the College Admission Test and its suitability as a diagnostic and decision-making tool for higher education institutions. Further, the convergence of content validation and factor analytic findings provides strong multi-method evidence of the test's validity. The expert-driven I-CVI established content relevance and representativeness, while EFA provided empirical support for the underlying factor structure, reinforcing the theoretical assumptions behind the test design. This dual validation process enhances the reliability, interpretability, and fairness of the College Admission Test, ensuring that it is both conceptually sound and statistically robust for use in academic placement and admission decisions.

3.4. Second pilot testing of the test for reliability analysis

As part of the 11-step development process, data from the second pilot test were analyzed to compute the reliability coefficient and norm. Reliability of tests is determined by the degree of congruence of results for different testing periods, which ensures that the test is stable and measures the same content each time.

The results shown in Table 4 indicate the reliability coefficients of subtests and the overall College Admission test. The overall reliability coefficient index of the test was 0.76, which is indicative of a moderately acceptable level of consistency. Among the major test domains, the Verbal Reasoning test recorded a high 0.80 reliability coefficient that indicates a significant level of internal consistency, proving it is accurate enough for an admissions test. The Vocabulary (0.73), as the subtests, also recorded a moderate and acceptable reliability level may be due to its more focused scope. Existing literature indicates that a reliability coefficient exceeding 0.70 is typically deemed acceptable for educational evaluations (Schumacker, 2005; Orongan, 2020), which implies that the results of this study corroborate the stability and consistency of the revised college admission test.

However, the relatively lower coefficient of Analogy (0.50) and Language Usage (0.56) suggests limited internal consistency. This could result from item heterogeneity, varying difficulty levels or ambiguous item formulations. While this result calls for item refinement to enhance measurement precision, DeVellis (2017) stated that values above 0.50 may still be deemed provisionally acceptable, particularly when item content diversity is intentionally broad in exploratory test development stages.

Meanwhile, the Numerical Ability test exhibited a moderate reliability of 0.65, thus making it an acceptable evaluation tool for similar purposes, but it needs further enhancement, specifically on either item format or item selection. While Post (2016) claimed that moderate correlation coefficients ranging from 0.40 to 0.60 do not provide adequate support for claims of reliability, Kline (2005) highlighted that Numerical subtests often exhibit variability in reliability due to a wide range of problem types and computational skills assessed.

The reliability estimates presented in this study affirm that the overall test is a moderately reliable instrument suitable for preliminary implementation in college admissions. However, further improvement is needed, particularly in the Analogy and Language Usage subtests. For future revision of this test, the need to enhance the items for clarity, difficulty balance, and construct alignment is highly recommended to increase internal consistency. To ensure that admitting incoming students is based on stable and consistent measures, the need to improve the reliability of lower-performing subtests must be taken into consideration.

This is particularly crucial given that reliability is an important element in psychometric assessments, as it ensures that variations observed in measurement outcomes are attributable to genuine differences among individuals rather than inconsistencies inherent to the assessment tool itself (Aldridge et al., 2017; Souza et al., 2017; Miller, 2019).

Table 4: Reliability coefficients for each test of the college admission test

Test	Number of Items	Reliability coefficient		Remarks
Verbal reasoning	65	0.80	High	Highly acceptable
Analogy	21	0.50	Moderate	Acceptable
Language usage	23	0.56	Moderate	Acceptable
Vocabulary	21	0.73	Moderate	Acceptable
Numerical ability	60	0.65	Moderate	Acceptable
Overall	125	0.76	Moderate	Acceptable

3.5. Finalization of the items of the college admission test

Table 5 presents the items removed from the Verbal Reasoning and Numerical Ability sections after conducting the two pilot tests. A total of five items were removed from Verbal Reasoning, while Numerical Ability retained all its items. This resulted in a final test composition of 120 items.

The limited number of discarded items from the Verbal Reasoning section demonstrates the overall soundness of the test's initial item construction. The removed items were identified based on psychometric criteria such as low item discrimination, extreme item difficulty, or dysfunctional distractors – factors that reduce quality and impact overall test validity. For example, removing only 1 out of 21 items in Analogy suggests high consistency among items in assessing analogical reasoning, despite this subtest showing a lower reliability coefficient in Table 4. Similarly, the removal of 3 items in Language Usage may indicate challenges in grammar or structure that can be corrected in future test iterations.

The removal of specific items ensures that only the most reliable and valid questions remain in the final version of the test. The refined test structure enhances its overall effectiveness in evaluating students' academic potential. The removal of these underperforming items is known to enhance the validity of the test as it focuses on test refinement and optimization (Zimmermann et al., 2017; Boateng et al., 2018).

Furthermore, researchers such as Lin (2018) and Haladyna and Rodriguez (2021) highlighted that the exclusion of ineffective items does not detract from the comprehensiveness of the test; instead, it enhances its measurement accuracy.

Interestingly, all items in the Numerical Ability test were retained after item analysis, indicating that the 60 items met the necessary standards of psychometric adequacy. This aligns with the test's acceptable reliability coefficient (0.65), as shown in Table 4, and implies that the quantitative reasoning items were generally well-constructed. However, the total absence of item removal may be interpreted with caution. While this could reflect good initial item quality, it also raises the possibility of leniency in evaluation criteria or a need for more stringent item analysis protocols. Future validation studies may also include differential item functioning analysis to examine potential item bias, especially in high-stakes admission contexts.

Consequently, this refinement process is expected to provide a more accurate and equitable measure of student admission for a higher education institution in the Cordillera Administrative Region, Philippines.

Table 5: Removed items from verbal reasoning and numerical ability

Test	Total numbers of removed items	Total items remained
Verbal reasoning	5	60
Analogy	1	20
Language usage	3	20
Vocabulary	1	20
Numerical ability	None	60
Total Items	-	120

3.6. Psychometric properties of the college admission test

3.6.1. Final content validity and reliability

Table 6 presents the final Item-Content Validity Index (I-CVI) after item removal and recalibration. The recalculated I-CVI of 0.93 indicates an increase of 0.06 from the initial I-CVI of 0.87. This improvement suggests that the revised version of the test better aligns with the intended constructs for measuring college readiness. The consistently high scores across all subtests affirm the test's robustness in measuring students' academic competencies.

Table 6 also presents the final reliability of the college admission test. The overall coefficient for the final version of the instrument is 0.75, interpreted as moderate but acceptable, particularly for a multi-dimensional instrument intended for education decision-making. This value suggests a satisfactory level of internal consistency in measuring general academic ability and aligns with standard benchmarks in educational assessment design (Tavakol and Dennick, 2011).

The Verbal Reasoning achieved an I-CVI of 0.93 and a reliability coefficient of 0.80, indicating both strong content coverage and high internal consistency. However, the subtests showed some variations. The Analogy subtest registered the lowest values (I-CVI = 0.80, Reliability = 0.51), suggesting that although the content is still acceptable, internal consistency remains limited. This may be due to the diverse cognitive operations involved in analogy reasoning or item heterogeneity – an observation also noted by Haladyna and Rodriguez (2013), who emphasized that analogy items can be challenging to standardize due to varying levels of abstraction and cultural influence.

On the other hand, Language Usage and Vocabulary achieved perfect I-CVI scores (1.0) and moderate reliability coefficients of 0.60 and 0.71, respectively. The perfect I-CVI scores suggest unanimous expert agreement on the content quality of these items, and their reliability scores indicate acceptable internal consistency for medium-stakes testing purposes. This reflects well-established patterns in test construction literature showing that language-based assessments, especially vocabulary, tend to have stronger psychometric performance due to their relatively constrained scope and direct instructional alignment.

Numerical Ability posted an I-CVI of 0.93 and a reliability coefficient of 0.65, both of which fall within acceptable ranges. While these results affirm the quality of the test content and its moderate reliability, they also suggest the need for ongoing refinement of numerical items to improve consistency across different types of quantitative problems (Kline, 2005).

The high values of I-CVI suggest that the test measures relevant verbal and numerical constructs, while the moderate to high reliability coefficient reflects that the test can produce consistent results, which is essential for decision-making in admissions. These findings support the argument that a localized, empirically validated admission test can offer a more equitable and accurate measure of student aptitude compared to standardized foreign-made tests. Furthermore, the psychometric properties demonstrated that the test met the criteria for standardized educational assessments, ensuring fairness and accuracy in evaluating students' competencies.

However, the relatively lower reliability of the Analogy subtest implies a need for continuous improvement. It is recommended that future test iterations could benefit from increasing the number of items, refining cognitive load, and pilot testing with broader demographic samples to improve item homogeneity. It is recommended that item response theory (IRT) modeling or differential item functioning (DIF) analysis be conducted as post-administration analyses to ensure fairness and precision.

Table 6: Final I-CVI and reliability of the college admission test

Tests	I-CVI	Reliability coefficient
Verbal reasoning	0.93	0.80
Analogy	0.80	0.51
Language usage	1.0	0.60
Vocabulary	1.0	0.71
Numerical ability	0.93	0.65
IFSU-CAT Over-All I-CVI and reliability	0.93	075

IFSU-CAT: Ifugao state university–college admission test; I-CVI: Item-content validity index

3.6.2. Concurrent validity results

Table 7 presents the concurrent validity of the newly developed College Admission Test was assessed by correlating its subtest scores with those from the OLSAT, which is widely recognized for measuring verbal and quantitative reasoning abilities. The analysis involved 1,700 incoming first-year students who took both assessments. The Verbal Reasoning was significantly correlated with the corresponding OLSAT verbal score, yielding a Pearson correlation coefficient of $r = 0.81$, $p < .0001$. Similarly, the Numerical Ability showed a strong positive correlation with the OLSAT numerical score, with a coefficient of $r = 0.89$, $p < .001$. These results suggest a high level of concurrent validity for both tests.

Concurrent validity is demonstrated when scores from a newly developed assessment correlate highly with an established instrument measuring the same construct (Cronbach and Meehl, 1955). The high correlation between the College Admission Test and the OLSAT indicates that the developed tests effectively capture similar constructs of verbal and quantitative reasoning, as measured by a well-established test. According to the American Educational Research Association, strong concurrent validity supports the test's interpretative appropriateness for use in educational decision-making.

The Verbal Reasoning correlation aligns with findings from McGrew (2009), who emphasized that verbal aptitude, when assessed through structured analogies and vocabulary-based reasoning, tends to produce consistent results across various cognitive batteries. This supports the conclusion that the

College Admission Test's verbal section possesses adequate construct alignment with the OLSAT. The even stronger Numerical Ability correlation reflects the robust overlap between the computation and reasoning processes assessed in both instruments.

These findings also resonate with research by Abma et al. (2016), who noted that concurrent validity correlations above 0.70 are considered strong evidence of convergent construct representation, especially in high-stakes cognitive testing. The statistical strength of these correlations indicates that the developed College Admission Tests perform comparably to the OLSAT in discriminating among individuals based on verbal and numerical reasoning abilities. The findings of the concurrent validity analysis carry important implications for both practice and policy in higher education admissions. The newly developed College Admission Test may provide meaningful insights into students' cognitive readiness for college-level work. The strong evidence of concurrent validity enhances institutional confidence in adopting the said test as a credible alternative to commercially standardized tests, particularly in contexts where such tools may lack cultural and curricular relevance.

The newly developed College Admission Test offers a locally developed solution that retains psychometric rigor while reflecting the competencies taught in the regional academic environment.

Table 7: Concurrent validity of the college admission test and OLSAT

Tests	N	r	p-value	Interpretation
Verbal reasoning	1,700	0.81	.001	Significant positive correlation (high concurrent validity)
Numerical ability	1,700	0.89	.001	Significant positive correlation (high concurrent validity)

3.6.3. Norms

The raw scores were transformed to allow comparison among examinees. The sample consisted of 169 students, which was considered adequate and representative of the Grade 12 population targeted by the test. A 120-item norm was developed for use in higher education institutions. For the validated admission test, the mean score was 48.62 with a standard deviation of 10.17. Raw scores were converted to z-scores, which were then scaled into the CEEB standard scores with a mean of 500 and a standard deviation of 100. This conversion was carried out by multiplying the z-score by 100 and adding 500.

In addition to CEEB scores, Deviation Intelligence Quotients (DIQ) were also applied. DIQ is another form of standard score with a mean of 100 and a fixed standard deviation of 15. The conversion was done by multiplying the z-score by 15 and adding 100.

Stanine scores, or standard nine scores, were also used to simplify interpretation. Raw scores were first converted to z-scores and then classified into whole-number categories from one to nine. Scores of -1.75 and below corresponded to stanine 1, between -1.75 and -1.25 to stanine 2, between -1.25 and -0.75 to stanine 3, between -0.75 and -0.25 to stanine 4, between -0.25 and 0.25 to stanine 5, between 0.25 and 0.75 to stanine 6, between 0.75 and 1.25 to stanine 7, between 1.25 and 1.75 to stanine 8, and 1.75 and above to stanine 9. Stanine scores from four to six are considered average, scores of three or lower are below average, and scores of seven or higher are above average.

4. Conclusion and recommendation

The developed and validated college admission test is an empirically based and reputable tool that may help incoming college students in a Philippine higher education institution in the Cordillera Administrative Region. This test underwent proper development and validation, making it an effective assessment tool. The study reaffirms that localized assessments provide a more accurate measure of student preparedness and align more closely with contextual academic demands. The implementation of this admission test may contribute to a more equitable selection process, addressing the limitations of foreign-made standardized tests and enhancing the fairness of college admissions.

It is recommended that the developed and validated college admission test for a higher education institution in the Philippines be adopted as a standardized tool for student selection. Regular validation and reliability testing of the instrument may be conducted. Local norms for gender, age, and course may be established to better interpret test scores.

Future studies may explore the predictive validity of test scores with academic outcomes. Finally, while the test is designed for the Cordillera region, its potential for national adaptation may be explored. Collaborations with educational bodies can help determine their applicability in other regions, allowing for necessary contextual modifications to address varying academic demands.

List of abbreviations

ACT	American college test
CEEB	College entrance examination board
DIQ	Deviation intelligence quotient
DIF	Differential item functioning
EFA	Exploratory factor analysis
I-CVI	Item-content validity index
IFSU-CAT	Ifugao state university–college admission test
IQ	Intelligence quotient
IRT	Item response theory
KMO	Kaiser–Meyer–Olkin
KR-20	Kuder–Richardson formula 20
NCEE	National college entrance examination
OLSAT	Otis-Lennon school ability test
SAT	Scholastic aptitude test
SD	Standard deviation

Funding

This study was funded by Ifugao State University through its Research and Development Department, which supported all stages from proposal to dissemination without influencing the research process or results.

Acknowledgment

The researcher gratefully acknowledges Ifugao State University, the participating schools, respondents, colleagues, and friends for their vital support in completing this study.

Compliance with ethical standards

Ethical considerations

The study followed institutional ethics, with informed consent, voluntary participation, secure data storage, and coded identifiers to ensure confidentiality.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abma IL, Rovers M, and van der Wees PJ (2016). Appraising convergent validity of patient-reported outcome measures in systematic reviews: Constructing hypotheses and interpreting outcomes. BMC Research Notes, 9: 226. https://doi.org/10.1186/s13104-016-2034-2 **PMid:27094345 PMCid:PMC4837507**

Aldridge V, Dovey T, and Wade A (2017). Assessing test-retest reliability of psychological measures: Persistent methodological problems. European Psychologist, 22(4): 207–218. https://doi.org/10.1027/1016-9040/a000298

Almanasreh E, Moles R, and Chen TF (2019). Evaluation of methods used for estimating content validity. Research in Social and Administrative Pharmacy, 15(2): 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066 **PMid:29606610**

Ashraf Z and Jaseem K (2020). Classical and modern methods in item analysis of test tools. International Journal of Research, 7(8): 397–403.

Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, and Young SL (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. Frontiers in Public Health, 6: 149. https://doi.org/10.3389/fpubh.2018.00149 **PMid:29942800 PMCid:PMC6004510**

Cerado EC and Garcia MA (2022). Development of continuous improvement program assessment tool (CIPAT) in the Department of Education-SOX Region, Philippines. Current Research in Language, Literature and Education, 7: 154–170. https://doi.org/10.9734/bpi/crlle/v7/16627D

Cronbach LJ and Meehl PE (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4): 281–302. https://doi.org/10.1037/h0040957 **PMid:13245896**

Dechavez CFJ (2024). Contextualizing learning: A multi-variable analysis of student characteristics, educational settings, and academic success. International Journal of Research and Scientific Innovation, 11(8): 228–243. https://doi.org/10.51244/IJRSI.2024.1108019

DeVellis RF (2017). Scale development: Theory and applications. 4th Edition, SAGE Publications, Thousand Oaks, USA.

Faul F, Erdfelder E, Buchner A, and Lang A-G (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41: 1149–1160. https://doi.org/10.3758/BRM.41.4.1149 **PMid:19897823**

Ferrando PJ and Morales-Vives F (2023). Is it quality, is it redundancy, or is model inadequacy? Some strategies for judging the appropriateness of high-discrimination items. Anales de Psicología, 39(3): 517–526. https://doi.org/10.6018/analesps.535781

Gatcho ARG, Manuel JPG, and Hajan BH (2024). No child left behind, literacy challenges ahead: A focus on the Philippines. Frontiers in Education, 9: 1349307. https://doi.org/10.3389/feduc.2024.1349307

Green JH, Davis C, Harmes M, Judith K, and Weideman A (2025). Using a five-phase applied linguistics design to develop a contextualized academic literacy placement test for pre-university pathway students. Literacy Research and Instruction, 64(2): 229-255. https://doi.org/10.1080/19388071.2024.2340031

Hair JF, Black WC, Babin BJ, and Anderson RE (2014). Multivariate data analysis. 7th Edition, Pearson New International Edition, Essex, UK.

Haladyna TM and Rodriguez MC (2013). Developing and validating test items. 1st Edition, Routledge, New York, USA. https://doi.org/10.4324/9780203850381

Haladyna TM and Rodriguez MC (2021). Using full-information item analysis to improve item quality. Educational Assessment, 26(3): 198–211. https://doi.org/10.1080/10627197.2021.1946390

Han K, Colarelli SM, and Weed NC (2019). Methodological and statistical advances in the consideration of cultural diversity in assessment: A critical review of group classification and measurement invariance testing. Psychological Assessment, 31(12): 1481–1496. https://doi.org/10.1037/pas0000731 **PMid:31763873**

Kaiser HF (1974). An index of factorial simplicity. Psychometrika, 39(1): 31–36. https://doi.org/10.1007/BF02291575

Kline TJB (2005). Psychological testing: A practical approach to design and evaluation. SAGE Publications, Thousand Oaks, USA. https://doi.org/10.4135/9781483385693

Kyriazos TA and Stalikas A (2018). Applied psychometrics: The steps of scale development and standardization process. Psychology, 9: 2531–2560. https://doi.org/10.4236/psych.2018.911145

Lee O, Santau AO, and Maerten-Rivera J (2011). Cultural validity in assessment: Addressing linguistic and cultural diversity. Routledge, New York, USA.

Lin C (2018). Effects of removing responses with likely random guessing under Rasch measurement on a multiple-choice language proficiency test. Language Assessment Quarterly, 15(4): 406–422. https://doi.org/10.1080/15434303.2018.1534237

Lynn MR (1986). Determination and quantification of content validity. Nursing Research, 35(6): 382–385. https://doi.org/10.1097/00006199-198611000-00017

Magno C and Gonzales RDLC (2011). Measurement and evaluation in the Philippine higher education: Trends and development. In the UNESCO Policy Series: Trends and Development in Philippine Education. UNESCO, Paris, France: 47–58.

Maruyama G, Ovies-Bocanegra MA, Do T, Peczuh MC, and Weisen S (2024). How much do we need college admission tests?

Analyses of Social Issues and Public Policy, 24: 1288–1308. https://doi.org/10.1111/asap.12417

McGrew KS (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. Intelligence, 37(1): 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Messick S (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9): 741–749. https://doi.org/10.1037//0003-066X.50.9.741

Miller MD (2019). Reliability in educational assessments. Oxford University Press, New York, USA. https://doi.org/10.1093/obo/9780199756810-0228

Nagy W, Townsend D, Lesaux NK, and Schmidt N (2012). Words as tools: Learning academic vocabulary as language acquisition. Reading Research Quarterly, 47: 91–108. https://doi.org/10.1002/RRQ.011

Orongan RC (2020). Reliability analysis on teachers' quarterly classroom assessment in basic education. Liceo Journal of Higher Education Research, 16(1): 99-107. https://doi.org/10.7828/ljher.v16i1.1370

Ozturgut O (2011). Learning by example: Standardized testing in the cases of China, Korea, Japan, and Taiwan. Academic Leadership: The Online Journal, 9(3): 13. https://doi.org/10.58809/DPYD4742

Park SK and Kim EG (2021). A study on the reliability and validity of the Korean health literacy instrument for late school-aged children. International Journal of Environmental Research and Public Health, 18(19): 10304. https://doi.org/10.3390/ijerph181910304 **PMid:34639605 PMCid:PMC8508180**

Pearce J, Edwards D, Fraillon J, Coates H, Canny BJ, and Wilkinson D (2015). The rationale for and use of assessment frameworks: Improving assessment and reporting quality in medical education. Perspectives on Medical Education, 4(3): 110–118. https://doi.org/10.1007/S40037-015-0182-Z **PMid:25962966 PMCid:PMC4456467**

Penn S (2023). Uses and abuses of standardised testing: Perceptions from high-performing, socially disadvantaged schools. Issues in Educational Research, 33(1): 266–283.

Polit DF and Beck CT (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing and Health, 29: 489–497. https://doi.org/10.1002/nur.20147 **PMid:16977646**

Polit DF, Beck CT, and Owen SV (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Research in Nursing and Health, 30: 459–467. https://doi.org/10.1002/nur.20199 **PMid:17654487**

Post M (2016). What to do with "moderate" reliability and validity coefficients? Archives of Physical Medicine and Rehabilitation, 97(7): 1051–1052. https://doi.org/10.1016/j.apmr.2016.04.001 **PMid:27095143**

Primi R, Ferrão ME, and Almeida LS (2010). Fluid intelligence as a predictor of learning: A longitudinal multilevel approach applied to math. Learning and Individual Differences, 20(5): 446–451. https://doi.org/10.1016/j.lindif.2010.05.001

Quaigrain K and Arhin A (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. Cogent Education, 4(1): 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Rudolph MJ, Daugherty KK, Ray ME, Shuford VP, Lebovitz L, and DiVall MV (2019). Best practices related to examination item construction and post-hoc review. American Journal of Pharmaceutical Education, 83(7): 7204. https://doi.org/10.5688/ajpe7204 **PMid:31619832 PMCid:PMC6788158**

Schumacker RE (2005). Standards for interpreting reliability coefficients. Joint Committee of the American Educational Research Association, Washington, D.C., USA.

Shi J, Mo X, and Sun Z (2012). Content validity index in scale development. Journal of Central South University: Medical Sciences, 37(2): 152–155.

Souza ACD, Alexandre NMC, and Guirardello EDB (2017). Psychometric properties in instruments evaluation of reliability and validity. Epidemiologia e Serviços de Saúde, 26(3): 649–659. https://doi.org/10.5123/S1679-49742017000300022 **PMid:28977189**

Sternberg RJ (2018). Context-sensitive cognitive and educational testing. Educational Psychology Review, 30(3): 857–884. https://doi.org/10.1007/s10648-017-9428-0

Tavakol M and Dennick R (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2(2): 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd **PMid:28029643 PMCid:PMC4205511**

Tavakol M and Doody G (2015). Making students' marks fair: Standard setting, assessment items and post hoc item analysis. International Journal of Medical Education, 6: 38–39. https://doi.org/10.5116/ijme.54e8.86df **PMid:25725229 PMCid:PMC4383637**

Tavakol M and Wetzel A (2020). Factor analysis: A means for theory and instrument development in support of construct validity. International Journal of Medical Education, 11: 245–247. https://doi.org/10.5116/ijme.5f96.0f4a **PMid:33170146 PMCid:PMC7883798**

van der Maas HLJ, Dolan CV, Grasman RP, Wicherts JM, Huizenga HM, and Raijmakers ME (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. Psychological Review, 113(4): 842–861. https://doi.org/10.1037/0033-295X.113.4.842 **PMid:17014305**

Yahia A (2022). Post-validation item analysis to assess the validity and reliability of multiple-choice questions at a medical college with an innovative curriculum. The National Medical Journal of India, 34(6): 359–362. https://doi.org/10.25259/NMJI_414_20 **PMid:35818102**

Yusoff MSB (2019). ABC of content validation and content validity index calculation. Education in Medicine Journal, 11(2): 49–54. https://doi.org/10.21315/eimj2019.11.2.6

Zamanzadeh V, Ghahramanian A, Rassouli M, Abbaszadeh A, Alavi-Majd H, and Nikanfar AR (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. Journal of Caring Sciences, 4(2): 165–178. https://doi.org/10.15171/jcs.2015.017 **PMid:26161370 PMCid:PMC4484991**

Zhou Y, Liu Q, Wu J, Wang F, Huang Z, Tong W, Xiong H, Chen E, and Ma J (2021). Modeling context-aware features for cognitive diagnosis in student learning. In the Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, Virtual Event, Singapore: 2420–2428. https://doi.org/10.1145/3447548.3467264 **PMid:34538113**

Zimmermann S, Klusmann D, and Hampe W (2017). Correcting the predictive validity of a selection test for the effect of indirect range restriction. BMC Medical Education, 17: 246. https://doi.org/10.1186/s12909-017-1070-5 **PMid:29228995 PMCid:PMC5725878**