2026, 30(1), 4–12. Received 10 May 2026. Accepted 1 June 2026.

STATISTICAL VALIDATION OF AN ENGLISH FOR SPECIFIC PURPOSES ACHIEVEMENT TEST USING CLASSICAL TEST THEORY: A CASE STUDY OF A HIGHER EDUCATION INSTITUTION

Mindaugas Kavaliauskas
Šiaulių valstybinė kolegija / Higher Education Institution

Annotation

Language achievement tests are widely used in higher education to assess student learning outcomes; however, their psychometric quality is rarely examined. This is particularly common in professional technical education, where teachers often design tests independently without conducting statistical validation. This article presents a Classical Test Theory (CTT)-based validation of a Professional English achievement test administered to first-year students in the Technical Maintenance of Automobiles programme at Šiaulių valstybinė kolegija, Lithuania. The study was motivated by a professional development partnership between the institution and Universidad de Jaén, Spain, within the NEOLAiA European Universities Alliance. A multiple-choice pilot test was analyzed using IBM SPSS Statistics. The analysis focused on item facility values, distractor functioning, item discrimination, and internal reliability. The results indicated that 80% of items fell within the acceptable facility value range, 45 out of 50 items demonstrated acceptable to excellent discrimination, distractors functioned effectively, and overall internal reliability was satisfactory (Cronbach's α = .783). The findings confirm that the test demonstrates adequate psychometric quality and highlight the practical value of CTT as an accessible tool for institutional test validation.

Key words: Classical Test Theory (CTT); language achievement test; test validation; English for Specific Purposes (ESP); item analysis; facility value; discrimination index; internal reliability; vocational education.

Introduction

Language tests are widely used in educational institutions to make decisions about student achievement; however, their psychometric quality is rarely analyzed. This is particularly evident in professional technical education contexts, where teachers often design tests independently without applying statistical test validation procedures (Cruz, 2024; Bachman, 1995). The rapid growth of language testing across Europe has not been accompanied by a corresponding development of quality assurance mechanisms, a concern highlighted by Cruz (2016b) in the context of European higher education.

In Lithuanian higher education institutions, achievement tests in English for Specific Purposes (ESP) are routinely administered at the conclusion of study courses. However, these assessments are not subjected to statistical validation procedures. Consequently, decisions regarding student progression are based on instruments whose reliability, validity, and item-level characteristics have not been empirically established. This issue appears to reflect a broader European trend. Cruz Trapero (2022), in a study of language professionals in the Canary Islands, found that none of the participants had received formal training in assessment during their university education, despite possessing extensive professional experience in language testing.

The present study originated from a professional development collaboration between Šiaulių valstybinė kolegija and Universidad de Jaén, Spain, both members of the NEOLAiA – European Universities Alliance, a network of emerging European universities committed to inclusion, democratic values, and innovation in higher education. Within the framework of this collaboration, the author received specialized training in statistical test validation from internationally recognized

2026, 30(1), 4–12. Received 10 May 2026. Accepted 1 June 2026.

experts, including Joaquín Cruz (Universidad de Jaén, Spain), author of seminal works in the field (Cruz, 2016a; Cruz, 2016b; Cruz, 2024). This engagement with systematic validation methodologies provided the primary impetus for the empirical analysis presented in this study.

Two principal theoretical frameworks underpin the statistical analysis of language tests: Classical Test Theory (CTT) and Item Response Theory (IRT). Although IRT has become increasingly prominent in contemporary language assessment research (Cruz, 2016a; Green, 2011), CTT remains widely applicable and particularly well-suited to the evaluation of achievement tests administered within institutional, course-based contexts (Bachman, 1995). Given that the present study focuses on an end-of-course language achievement test administered to a relatively small cohort, CTT was deemed the most appropriate analytical framework.

Purpose and relevance of the article: The aim of this article is to present the statistical validation of an English language achievement test administered to first-year students enrolled in the Technical Maintenance of Automobiles programme at Šiaulių valstybinė kolegija, employing the framework of Classical Test Theory (CTT). The central research question guiding the study is as follows: to what extent does the administered English language test demonstrate acceptable psychometric properties when analyzed within the CTT paradigm? The subsequent sections of the article outline the theoretical framework, describe the research methodology, present the empirical results, and discuss the findings in relation to this research question.

Theoretical Framework

The Importance of Test Validation

Test validation refers to the systematic process of gathering and evaluating evidence to support the interpretation and intended use of test scores. Crucially, validity is not an inherent property of a test instrument; rather, it pertains to the appropriateness of the inferences drawn from test results and the decisions based upon them. As Lyle F. Bachman (1995, p. 237) argues, “the inferences regarding specific uses of a test are validated, not the test itself.” Accordingly, a given test score may be considered valid for one interpretative purpose, while being unsuitable or invalid for another.

Three main types of evidence support test validity: evidence that the test adequately represents the target content domain, evidence that test scores correlate with other measures of the same ability, and evidence that scores accurately reflect the underlying construct being assessed (Bachman, 1995). Among these, construct validity is particularly important, as it addresses the fundamental question of whether a test truly measures the construct it is intended to measure.

Test scores are often used to make important decisions about students, such as assigning grades or evaluating academic progression. Consequently, poorly constructed tests may lead to unfair results and inaccurate outcomes. As Bachman (1995) argues, the greater the consequences associated with test-based decisions, the stronger the need to ensure that test scores are both reliable and valid.

There is a common assumption that concerns about validity apply primarily to large-scale, formal proficiency examinations. While proficiency tests are designed to measure general language ability, achievement tests assess how well students

2026, 30(1), 4–12. Received 10 May 2026. Accepted 1 June 2026.

have learned the content of a particular course (Bachman, 1995). As a result, it may appear sufficient for an achievement test simply to reflect the material covered in the course. However, Bachman (1995, p. 291) emphasizes that "even achievement tests must undergo construct validation if their results are to be interpreted as indicators of ability." Content coverage alone cannot determine whether test items are appropriately constructed, whether distractors function effectively, or whether the test can distinguish between higher- and lower-performing students.

This issue is particularly relevant in institutions where teachers design tests without formal training in psychometrics. Cruz (2024) observes that most language teachers in European educational settings have not studied quantitative test analysis. This observation is supported empirically by Cruz Trapero (2022), whose survey of language professionals in the Canary Islands found that none of the participants had received specific university-level training in assessment at universities or higher education institutions, despite being actively involved in language test development. Yet the statistical tools available - item difficulty, discrimination indices, and reliability coefficients - are practical and accessible, and can be applied by any motivated teacher or researcher (Cruz, 2024; Green, 2011). As Cruz Trapero (2022, p. 135) concludes, training in assessment can serve "a double objective: creating valid assessment tools and improving the assessment skills of a whole community of language professionals." The present study addresses this gap within the context of Lithuanian higher education.

Classical Test Theory: Foundational Model and Core Analytical Parameters

Classical Test Theory (CTT) is the oldest and most widely used approach to the statistical analysis of test data (Bachman, 1995; Cruz, 2016a). Its origins can be traced to the early twentieth century, with Ch. Spearman's 1904 work on measurement and correlation commonly regarded as the foundation of the theory (Traub, 1997, as cited in Cruz, 2024). CTT has remained widely applied, because it yields clear, easily interpretable results that are directly useful for test development and improvement (Cruz, 2024; Green, 2011). Its suitability for teacher-designed tests in institutional settings has been demonstrated across a range of educational contexts, including studies validating multiple-choice tests developed by teachers at both school and higher-education levels (Gyamfi, 2022).

CTT is based on the fundamental assumption that a student's observed test score is not a perfectly accurate reflection of their true ability. According to the theory, every test score consists of two components: a true score, representing the student's actual level of knowledge or ability, and an error component arising from factors unrelated to that ability, such as anxiety, ambiguous test items, or random chance (Bachman, 1995). This relationship is expressed as: $X = T + E$, where X denotes the observed score, T is the true score, and E is the random measurement error (Bachman, 1995; Cruz, 2024; Gyamfi, 2022). Because true scores cannot be observed directly, the primary aim of CTT analysis is to minimise measurement error and ensure that observed scores reflect genuine differences in student ability as accurately as possible.

Four statistical parameters are employed in the present analysis. The first is the **facility value**, which indicates whether test items are of an appropriate level of difficulty. This tells us what percentage of students answered an item correctly. The higher the value, the easier the item (Green, 2011). In the case of multiple-choice

2026, 30(1), 4–12. Received 10 May 2026. Accepted 1 June 2026.

items, values between 20% and 80% are generally regarded as acceptable (Green, 2011; Cruz, 2024). In achievement tests, however, higher facility values are both expected and appropriate, as students have already been taught the tested material (Green, 2011).

Distractor analysis examines whether incorrect answer options in multiple-choice items are sufficiently plausible. For distractors to be considered functional, each should attract at least 7% of student responses (Cruz, 2024; Green, 2011). Distractors selected by fewer than 7% of test takers are generally regarded as implausible and should therefore be revised or replaced. Gyamfi (2022) identified non-functional distractors as one of the most common weaknesses in teacher-designed tests, a finding that is consistent with the results of the present study.

The discrimination index evaluates whether test items effectively distinguish between stronger and weaker students. In SPSS, this measure is reported as the Corrected Item Total Correlation (CITC), which reflects the relationship between a student's performance on an individual item and their performance on the test (Green, 2011). According to Cruz (2024), CITC values between 0.25 and 0.35 are considered acceptable, values between 0.36 and 0.45 are regarded as good, and values above 0.46 indicate excellent discrimination. Values below 0.25 suggest that an item discriminates poorly, whereas negative values may indicate more serious problems, such as an incorrectly keyed answer (Green, 2011).

Internal reliability, commonly measured using Cronbach's α, evaluates whether all test items function together to assess the same underlying ability. Cronbach's alpha values range from 0 to 1, with values above 0.70 generally regarded as the minimum acceptable standard for language tests, while values exceeding 0.80 are considered preferable (Green, 2011). Achievement tests often yield slightly lower alpha values than proficiency tests because post-instruction groups tend to be more homogeneous in ability (Bachman, 1995; Green, 2011). For comparison, Gyamfi (2022) reported a reliability coefficient of 0.73 for a comparable teacher-designed multiple-choice test, suggesting that values within this range are typical for achievement testing contexts.

CTT versus IRT: Rationale for the Choice

Two major analytical approaches are commonly distinguished in language testing: Classical Test Theory (CTT) and Item Response Theory (IRT). In recent decades, IRT has become increasingly prominent in large-scale language testing research because it provides sample-independent item parameter estimates; in other words, item statistics remain stable across different groups of test takers (Cruz, 2016a). However, IRT also involves several important practical limitations. Green (2011) notes that IRT analyses generally require a minimum sample size of approximately 200 participants, with 300 or more considered preferable, a condition that is often unattainable for institutional test developers. Furthermore, IRT relies on a logarithmic measurement scale that may be unfamiliar to many language teachers, and its software outputs are considerably more complex to interpret than those generated by IBM SPSS *Statistics* (Green, 2011; Cruz, 2024).

For institutional achievement tests administered to relatively small and homogeneous groups — such as in the present study, where — CTT represents a fully appropriate and methodologically justified approach (Bachman, 1995). In addition, CTT analyses can be conducted using widely accessible software and produce results that are readily interpretable even by non-specialist test developers

2026, 30(1), 4–12. Received 10 May 2026. Accepted 1 June 2026.

(Cruz, 2024). For these reasons, CTT was selected as the analytical framework for the present study.

Methodology

Institutional Context

Šiaulių valstybinė kolegija is a Lithuanian state higher education institution offering study programmes in the field of technology and health sciences. A ten-point criterion-referenced grading scale is applied across all subjects. Student assessment is cumulative, with the final grade calculated as the weighted combination of interim assessment results and the end-of-semester examination score. Across all study programmes and student groups, the Professional English examination is administered in the form of a written test.

Despite the widespread use of these examinations, no statistical validation of their quality has previously been conducted at Šiaulių valstybinė kolegija. Test design and administration are carried out independently by teachers, typically without formal psychometric training or statistical analysis to determine whether the tests accurately measure intended student learning outcomes. This situation reflects patterns identified in comparable European educational contexts, where language test developers frequently lack formal training in assessment (Cruz Trapero, 2022; Cruz, 2024). The decision to undertake the present validation study was motivated directly by the author's participation in the Language Test Design and Validation course at Universidad de Jaén, Spain, facilitated through the NEOLAiA European Universities Alliance.

The Test

To apply CTT analysis within this institutional context, a pilot Professional English achievement test was developed and administered. In the present study, the term *pilot* refers to a trial administration intended to generate item-level statistical data before the test is used for official grading purposes. The test was aligned with the syllabus of the Technical Maintenance of Automobiles study programme and consisted of 50 multiple-choice items, each containing one correct answer and two distractors. The multiple-choice format was selected because it enables objective scoring and is particularly suitable for CTT-based item analysis (Green, 2011; Cruz, 2024; Gyamfi, 2022). The test was administered at the end of the academic semester.

Participants

The test was administered to a group of 14 first-year students enrolled in the Technical Maintenance of Automobiles study programme. As typical in end-of-course achievement testing contexts, the group was relatively homogeneous in terms of language ability, since all participants had completed the same course under identical instructional conditions. Such homogeneity was expected to result in a restricted score range, which may in turn produce somewhat lower discrimination and reliability indices than those typically observed in more heterogeneous populations (Bachman, 1995; Green, 2011). This limitation was taken into consideration in the interpretation of the results.

2026, 30(1), 4–12. Received 10 May 2026. Accepted 1 June 2026.

Analytical Procedure

Student responses were entered into IBM SPSS *Statistics* and analyzed using standard procedures derived from CTT, following the methodological guidelines outlined in Cruz (2024) and Green (2011). The analysis focused on five areas: descriptive statistics, including the mean score and standard deviation; facility values for individual items; distractor analysis to evaluate the effectiveness of incorrect response options; item-total correlations as indicators of item discrimination; and Cronbach's alpha as a measure of internal reliability.

Results

The following results address the central research question of the study: to what extent does the test demonstrate acceptable psychometric properties when analyzed within the framework of CTT?

Descriptive Statistics

The group achieved a mean score of 74.3%, with a standard deviation of 9.8, indicating that most students performed satisfactorily after completing the Professional English course. The score distribution was approximately normal, although slightly positively skewed, suggesting that a greater proportion of students achieved scores toward the upper end of the scale. This pattern is consistent with the characteristics of post-instruction achievement testing contexts, in which most students have already studied the assessed material (Green, 2011).

Facility Values

The facility values for the 50 test items ranged from 0.38 to 0.87. Most items (80%) fell within the generally accepted range of 0.30-0.80 (Cruz, 2024; Green, 2011), indicating an appropriate overall level of difficulty. A small number of items produced facility values above 0.80, which is unsurprising in an end-of-course achievement test where students had already been exposed to the relevant material. Importantly, no item produced a facility value below 0.20, suggesting that none of the items was excessively difficult for this group of students. Overall, these findings are broadly consistent with results reported in comparable studies of teacher-designed tests (Gyamfi, 2022).

Discrimination Index

Item–total correlation values ranged from 0.22 to 0.58. Most items (45 out of 50) demonstrated acceptable to excellent discrimination indices, indicating that these items successfully differentiated between stronger and weaker test takers (Cruz, 2024; Green, 2011). Five items produced values slightly below the recommended 0.25 threshold and were therefore identified as candidates for revision in future versions of the test. No item showed a negative discrimination value, which might otherwise have indicated problems such as an incorrectly keyed answer or flawed item construction.

Distractor Analysis

Distractor analysis demonstrated that most incorrect response options attracted a sufficient proportion of student responses across the test items (Cruz, 2024). However, two distractors across four items were selected by fewer than 7% of test takers, suggesting that these options were too implausible to function effectively. These distractors should therefore be revised or replaced in future administrations of the test. Non-functional distractors are commonly reported in studies of teacher-

2026, 30(1), 4–12. Received 10 May 2026. Accepted 1 June 2026.

designed multiple-choice tests (Gyamfi, 2022) and represent one of the most practically useful outcomes of Classical Test Theory analysis.

Internal Reliability

The overall Cronbach's α coefficient was .783, exceeding the minimum acceptable threshold of .70 commonly recommended for language tests (Green, 2011; Bachman, 1995). This result indicates satisfactory internal consistency, suggesting that the test items function cohesively to measure the same underlying construct. The "Cronbach's Alpha if Item Deleted" output in IBM SPSS *Statistics* further demonstrated that removing any individual item would not substantially affect the overall α coefficient, with the exception of one item whose removal would increase α slightly to .801.

Achievement tests typically yield lower reliability coefficients than proficiency tests because students within post-instruction groups tend to display relatively similar ability levels, thereby reducing score variability (Green, 2011; Bachman, 1995). Within this context, a coefficient of .783 may therefore be regarded as a strong result. It also compares favorably with the coefficient of 0.73 reported by Gyamfi (2022) in a comparable study of a teacher-designed multiple-choice test.

Discussion

Taken together, the results of the CTT analysis indicate that the Professional English achievement test administered at Šiaulių valstybinė kolegija demonstrates satisfactory psychometric quality. The facility values, discrimination indices, distractor functioning, and internal reliability coefficient all fall within, or close to, the ranges recommended in the literature (Cruz, 2024; Green, 2011; Bachman, 1995). This constitutes a meaningful finding, as it provides the first empirically grounded evaluation of test quality at the institution and confirms that the test measures student learning outcomes with an acceptable degree of accuracy. The reliability coefficient of .783 also compares favorably with results reported in comparable studies of teacher-designed tests (Gyamfi, 2022), suggesting that experienced teachers may be capable of developing adequately functioning assessment instruments even without formal psychometric training.

At the same time, the analysis identified several areas requiring further improvement. Five items with discrimination values below 0.25 and two non-functioning distractors across four items constitute clear targets for revision. Importantly, these weaknesses would likely have remained undetected without statistical analysis (Cruz, 2024). This highlights the diagnostic value of CTT: rather than merely confirming or rejecting a test, it enables test developers to identify specific and actionable areas for improvement. In addition, the item flagged in the "Cronbach's Alpha if Item Deleted" analysis — whose removal would increase the overall reliability coefficient to .801 — should be reviewed carefully to determine whether it may be measuring a construct different from that assessed by the remainder of the test.

The absence of prior statistical validation at the institution reflects a broader pattern across European educational contexts, where language testing is frequently treated as an administrative procedure rather than as a measurement activity requiring empirical verification (Cruz, 2016b; Cruz, 2024). Cruz Trapero (2022) demonstrated in the Spanish context that even experienced language professionals often receive no formal university-level training in assessment, and that structured training initiatives can significantly improve both the quality of assessment

2026, 30(1), 4–12. Received 10 May 2026. Accepted 1 June 2026.

instruments and the professional competence of teachers. The present study demonstrates that CTT-based validation is both feasible and informative within a Lithuanian higher education context, requiring no specialized software beyond IBM SPSS *Statistics* and no advanced statistical training.

Conclusions

This study set out to investigate the extent to which the English language achievement test administered to students of the Technical Maintenance of Automobiles study programme at Šiaulių valstybinė kolegija demonstrates acceptable psychometric properties when analyzed within the framework of CTT. The findings - including acceptable facility values, predominantly positive item discrimination indices, largely functional distractors, and a Cronbach's α coefficient of .783 - support the conclusion that the test constitutes a psychometrically sound instrument for measuring end-of-course English language achievement.

As Cruz Trapero (2022) demonstrated in comparable professional contexts, training in assessment can simultaneously enhance the quality of measurement instruments and improve the assessment literacy of practitioners. The present study further shows that, by using widely accessible tools and relatively straightforward analytical procedures, it is possible to generate meaningful empirical evidence regarding test quality, with direct practical implications for test development, revision, and implementation.

Several limitations of the present study should be acknowledged. First, the sample size of 14 students is relatively small, and the single-group, single-test design limits the extent to which the findings can be generalized beyond the specific context under investigation. Furthermore, the study does not aim to provide a comprehensive validity argument and therefore makes no broader claims regarding construct validity. Future research could apply the same CTT methodology across multiple student groups and academic specializations. In addition, analyses based on IRT, which typically require a minimum sample of approximately 200 participants (Green, 2011), would constitute a logical next step for studies involving larger student populations.

Recommendations

Drawing on the results of the present analysis, several recommendations can be made for higher education institutions and language professors:

- Higher education institutions should establish minimum quality assurance requirements, including the application of Classical Test Theory (CTT) analysis to all end-of-course language achievement tests prior to their use in grading and progression decisions.

- Widely available statistical software such as IBM SPSS Statistics enables the efficient implementation of CTT procedures and provides clear, interpretable outputs; thus, advanced psychometric software is not required for conducting basic validation analyses.

- Test items identified as problematic through discrimination indices or distractor analysis should be systematically revised and improved, rather than discarded or reused without modification.

2026, 30(1), 4–12. Received 10 May 2026. Accepted 1 June 2026.

- Institutional collaboration initiatives, such as participation in the NEOLAiA – European Universities Alliance, alongside targeted training programmes delivered by experts including Joaquín Cruz (Universidad de Jaén), should be expanded to a broader cohort of academic staff. Empirical evidence (e.g., Cruz Trapero, 2022) indicates that structured training in assessment contributes to measurable improvements in both test quality and teacher competence.
- Given that all study programmes at Šiaulių valstybinė kolegija incorporate end-of-course language testing, the methodological approach demonstrated in this study should be systematically extended across different specializations, thereby enabling comparative analysis between student groups.

- Future research should incorporate Item Response Theory (IRT) analyses to compare findings with CTT-based results and to explore potential variation in achievement test design across disciplinary contexts.

References

Bachman, L. F. (1995). Fundamental considerations in language testing. Oxford University Press.

Cruz, J. (2016a). A protocol to design a CEFR-linked proficiency rating scale for oral production and its app implementation [Doctoral dissertation, Universidad de Jaén].

Cruz, J. (2016b). The industry of testing and mutual recognition in Southern Europe. The Grove: Working Papers on English Studies, 23, 11–22.

Cruz, J. (2024). Diseño y validación de exámenes de dominio de lengua. Ministerio de Educación, Formación Profesional y Deportes.

Cruz Trapero, J. (2022). Assessment literacy through the design of analytic scales. In S. Hidri (Ed.), *Language Assessment Literacy and Competence: Volume 2* (pp. 127–137). Cambridge University Press.

Green, R. (2011). Statistical analysis for language testers. Palgrave Macmillan.

Gyamfi, A. (2022). Application of Classical Test Theory (CTT) in the validation of teacher made Mathematics Multiple Choice Test (MMCT) items. *Asian Journal of Advanced Research and Reports, 16*(11), 1–12. https://doi.org/10.9734/AJARR/2022/v16i11434

Claude by Anthropic. Refinement. https://www.anthropic.com/claude.