



Exploring Construct Validity and Construct Underrepresentation in CBT-Based Use of English among Diploma Students.

Ali Mohammed Also¹, Rasheedah Suleiman², Abdullahi Kaigama³

¹ Federal Polytechnic, Damaturu, Yobe State, Nigeria. Email Address: alidaada2000@gmail.com

² Federal Polytechnic, Damaturu, Yobe State, Nigeria. Email: rasheedaalso@gmail.com

³ Federal Polytechnic, Damaturu, Yobe State, Nigeria Email: bbabdullahi@gmail.com

ABSTRACT

This study assesses construct validity and construct underrepresentation in Computer-Based Testing (CBT) for Use of English courses among diploma students. Whereas CBT improves evaluation scalability and grading efficacy, automated components mostly overstress low-level recollection via multiple-choice set of questions, unconsciously eliminating higher-order language skills like impulsive writing, practical competence, and discourse management. Applying a mixed-methods paradigm to analyse test items, students' feedback and performance metrics, the study highlights how procedural limitations and constricting item strategy lead to construct underrepresentation. Finally, the findings assert for the integration of hybrid item formats and genuine communicative tasks into Computer-based models to affirm detailed, valid assessments of academic English proficiency.

Keywords: *Construct Validity, Construct Underrepresentation, Computer-Based Testing, Diploma Students.*

Introduction

In educational and language assessment, testing serves as an operative mechanism via which abstract psychological items are transformed into quantifiable students' performance (Adebola, 2022). An evaluation instrument may serve as an evaluative model projected to infer a student's essential skills, knowledge and other operative abilities within a specific entity. Conversely, the implications drawn from test scores is reliant heavily on how precisely the evaluation is in consonance with the recognized curriculum values and perceptive frameworks (Anderson, 2018). In communicative language testing, high-value assessment needs a clear balance between receptive tasks which include listening and reading as well as productive skills comprising of writing and speaking. When an evaluation depends wholly on limited test designs due to organizational constraints, the test risks lessening the target item, which leads to invalid score explanations and other ambiguous indicators of student skill (Ariza-Quñones et al., 2022).

At the basic level, test items may represent the main building blocks of an evaluation, which acts as single prompts planned to obtain observable responses that may reflect particular linguistic or cognitive skills (Benton & Young, 2018). The operational superiority of a test item directly dictates its psychometric integrity. Well-formed test items must conform with the strict technical specifications, to maintain item's autonomy, to avoid direct textbook phrasing, and use reasonably functional distractors with dissimilar lower-performing students from higher-performing ones (Berk, 2018). Additionally, faulty item production, unclear phrasing, or grammatical prompts familiarize construct-irrelevant noise, averting the element from precisely apprehending the candidate's true aptitude level (Anderson, 2018).

In computerized testing domains including Computer-Based Testing (CBT), the correlations between test elements and construct validity become mainly sensitive. Computerized platforms frequently implement a heavy dependence on chosen-response element formats to allow for immediate, simulated scoring (Al-Mekhlafi, & Naji, 2013). Whereas this operative structure exploits scoring independence and logistical efficacy, it limits the sorts of test elements that can be used. Consequently, tricky communicative tasks that need practical discourse creation, impulsive interface, or conformation procedure are often omitted from the vigorous element pool. Assessing test element via response models, mental tenet coding, and distractor analyses is still vital that automated item banks to retain high construct validity, adequate item difficulty spread, and robust analytical efficacy (Ariza-Quñones et al., 2022).

Issue in focus

The transition from classical paper-based testing to Computer-Based Testing (CBT) across Nigerian higher institutions has remarkably streamlined examination organization, reduced grading turnaround periods and lessened instances of physical examination malpractice (Bartolome & Navalta, 2025). Conversely, this innovative shift brings with it severe validity threats when deployed to general language evaluation which include several hierarchical course under the component of Use of English for both National and Higher National diploma students. According to the prominent

Messick's integrated theory of construct validity, an assessment item is valid when it sufficiently specifies the hypothetical construct it attempts to assess without introducing peripheral variables. Within the purview of polytechnic education, where the Use of English curriculum is clearly planned around communicative language teaching (CLT) to develop both productive and receptive competencies, the CBT transcript only relies heavily on multi-choice (objective) elements for easy and automated scoring (Benton & Ryalls, 2016). The core productive skills such as mechanics of composition, essay writing and oral communication are exclusively excluded from the test bank and other procedures. This operational oversight results in an imperfect demonstration of language aptitude within the test bank and this might pose serious problem towards attaining effective assessment result. So, examining construct validity and construct under representation is thus, important to ensure that technical test formats deliver detailed, valid and contextually defensible assessment of students' attainment.

Purpose of the Study

The main aim of our study is to identify and assess the degree of construct underrepresentation in CBT-Based Use of English evaluations by comparing test bank item distributions in Diploma syllabus. Similarly, the study attempts to examine the construct validity of CBT-Based use of English test through the analysis of students' score discrepancies across CBT multiple-choice components and direct productive language tasks.

The Construct Computer-Based Testing

Computer-Based Testing (CBT) serve as a main innovative shift in educational evaluation to transform how examinations are advanced, administered and recorded across secondary and higher education domains. The general adoption of CBT is basically driven by its operative efficacy, enhanced examination security, automated grading abilities and swift score processing (Pengelley et al., 2025). By applying randomized item banks and safe digital domains, computer-based evaluations, essentially lessen security breaches, and physical administrative errors in comparison with the classical paper-and-pencil assessments (OECD, 2023). Additionally, CBT entities allow basic test developers to include multimedia items and flexible interaction features to enable more interactive and standardized testing situations across varied students' population (Botaccio et al., 2020).

Notwithstanding these operative motivations, the transition to CBT introduces crucial validity and justice trials that may weaken test score elucidation. When high-stakes language evaluations transition of digital methods, automated constraints frequently force a dependence on some selected-response multiple-choice items, methodically eliminating productive communication abilities like all-inclusive essay writing and oral interface (Pengelley et al., 2025; Tsagari & Bøhn, 2025). Furthermore, students' presentation can be mystified by construct-irrelevant elements including different stages of digital literacy, scree-inclined reading distress interaction navigation dormancy, and platform-based procedural apprehension (Abate & Mishore, 2024). Without vigorous established infrastructure, reasonable device admittance, and arduous psychometric endorsement to confirm mode comparability, CBT platforms threat determining students' procedural awareness rather than their accurate purview ability.

The Term Construct Validity

Construct validity signifies the introductory model for assessing whether an evaluation accurately measures the theoretic construct trait it is projected to assess. As expressed in Messick's landmark integrated model, construct validity assesses the extent to which experimental evidence support the sufficiency, appropriateness and social outcomes of score-inclined elucidations (Messick, 1989). Under this integrated pattern, authentication is not a fixed element to a test item itself, but an ongoing evaluative procedure that examines whether extrapolations drawn from test scores are systematically justified across precise administrative domains (Kane, 2013). It refers to the degree to which an item measures the construct that it intends to assess (Alshurafat et al., 2024). who presented this concept, projected that the construct validity of a specific tool should be appraised in the absence of a reliable accepted standard that is sufficient to define the construct determine to measure. It evaluates whether the function of the new tool is in line with the basic predictions made according to a specific theory (Chapelle, & Lee, 2021).

To produce a construct validity, assessment experts examine two key validity threats that can systematically undermine evaluation accuracy, which include construct-irrelevant variance and construct under representation (Febriyana & Harjanto, 2023). Construct-irrelevant occurs when an assessment establishes extraneous factors that may artificially inflate or deflate scores, like computer anxiety, difficult digital interaction or the mechanical navigation. Conversely, under representation takes place when a test fails to address essential dimensions of the targeted construct, and hence, render the assessment totally narrow and skew. Real validation needs experimental research that can clearly minimize both threats to affirm score variations reflect genuine variations in the targeted underlying trait.

In contemporary educational assessment, authentication study continuously operationalizes construct validity via argument-inclined authentication models (Khoy, 2025). An argument-inclined method needs researchers to present an explicit inference and some assumptions connecting a student's raw test performance to the last decision-making. Scholars then methodically test these basic assumptions through the collection of mental processes as well as washback studies. Ultimately, appraising construct validity affirms that high-stakes educational evaluations deliver precise, fair and contextually reliable measurements of students' ability. Construct validity is a introductory construct in social science study, affirming that experimental measures precisely capture the methodological constructs they are intended to characterize (Cronbach & Meehl, 1955).

The Aspect of Construct Underrepresentation

Construct underrepresentation forms one of the major threats to educational assessment and test score validity, which occur when an evaluation fails to sufficiently sample the full scope of the methodological domain it attempts to measure (Leung, 2025). In language testing, construct underrepresentation often show when evaluations emphasizes on lower-order recognition tasks or genuine, productive language performance. Contemporary communicative language models view language skill as a multidimensional construct comprising discourse structure, grammatical competence, sociolinguistic suitability and speaking abilities or productive writing. The practical and theoretic outcome of construct underrepresentation can be extended beyond imprecise score interpretations to strictly impact classroom instruction via negative washback. When an evaluation underrepresents its target curriculum, students and moderators align learning and teaching plans entirely with the constricted skill needed by the examination.

Construct underrepresentation (CU) has emerged as an essential model in language studies. It is a situation where the operationalization of a theoretic construct is incomplete, and vital components of that construct are not sufficiently engaged in evaluation, instructive practice or research (Gupta et al., 2025). When the operative definition of a construct fails to attain to the genuine scope, the consequences about language skill, proficiency or learning outcomes may be analytically inadequate and misleading (Menon et al., 2025). For example, studies on the TOEFL iBT speaking test has exhibited that the test format leaves no space for reciprocal interface, thus failing to analyse students' oral questioning aptitude. The study again found that instructors approved the essence of querying but tended not to be active on students' questioning skill's absence from the real test insufficient preparation time, and professional reputation concerns tied to test results. This demonstrates how CU functions not just as a procedural assessment problem but as a universal force that outlines instructive importance and possibly confines the expansion of reliable communicative capability (Sportel et al., 2023). Methodologically, construct underrepresentation presents challenges for the expansion and validation of instruments projected to measure tricky language learning constructs including self-efficacy, motivation and interactional aspect. The model of construct underrepresentation provides a potent and cogent lens for questioning whether present language tests, pedagogies and curricula sufficiently represent the full range of communicative competencies that students really require for academic, social and professional participation (Schneider, 2017).

Convergence Validity in Computer-Based Testing

Convergent validity signifies a central subtype of construct validity that assesses the extent to which two different assessment items that methodically target similar underlying construct produce reliable, highly correlated outcomes (Shou et al., 2022). Entrenched in the introductory multi-trait-multi-method (MTMM) matrix approach, convergent validity drives on the psychometric evidence that if an evaluation correctly captures a particular theoretic attribute, its scores must be in consonance with strong established and independent items of that same entity. In language and educational testing, establishing convergent validity includes the calculation of bivariate relationship coefficients between a newly executed testing instrument and validated external criteria. A numerically essential potent positive relationship provides experiential proof that the new evaluation successfully moves into the directed construct rather than apprehending random error or inessential variables (Menon, et al., 2025).

In the field of computer-inclined language testing, assessing convergent validity is necessary to confirm that digital or technical testing formats measure real language proficiency rather than procedural test-completeness. For example, when authenticating a technical Computer-Based Testing (CBT) *Use of English* component, academics assess convergent validity by associating students' CBT scores with their original performance on direct, human-assessed criterion tasks, such as homogeneous written conformations or oral interviews (Xie, 2024). If students' outcome on the CBT chosen-response elements show a powerful numerical correlation with direct productive language tasks, the CBT instrument shows potent convergent validity. Contrariwise, a weak relationship designates that the computerized format may suffer from construct-irrelevant variance, indicating that the digital assessment fails to satisfactorily reflect dependable, practical language abilities (Chapelle & Lee, 2021).

Approach and Design

Our study examined some sets of past examination questions to assess the construct validity and underrepresentation. Specifically, we employed a robust Documentary Content Analysis through a Deductive Coding Framework reliant on Alignment Matrices. This approach methodologically maps every test element from past CBT examination papers against the prescribed course syllabus aims and some recognized tenet like the Bloom's Improved Taxonomy (Sireci, 1998). By classifying each problem according to its directed language sub-skill, for instance, reading comprehension, grammar recognition, mechanics, and cognitive depth including lower-order recall and higher-order production). This paradigm permits the researchers to compute tangible percentage distributions across language domains, directly revealing where the test bank over-represents positive element types whereas entirely neglecting others (Bowen, 2023).

We initially applied for the release of two sets of Use of English Examinations Corpus for the analysis. As part of the exclusion criteria, we included only examination items that took place in 2024/2025 academic sessions. This gave us a sense of recency and complete immersion of students in the Computer-based Examinations. For the Exclusion criteria, we included only items that were related to the Use of English where HND examinations were excluded. The extraction of these constructs took place in two days and the task took place manually, without the employment of any software. Our team analyzed 83 different examination items with series of numerous multiple-choice questions. No essay was found in the examination item. A set of unique code were assigned by the researchers to assist in the process of analysis and the following outcome were realized.

Analysis and Discussion

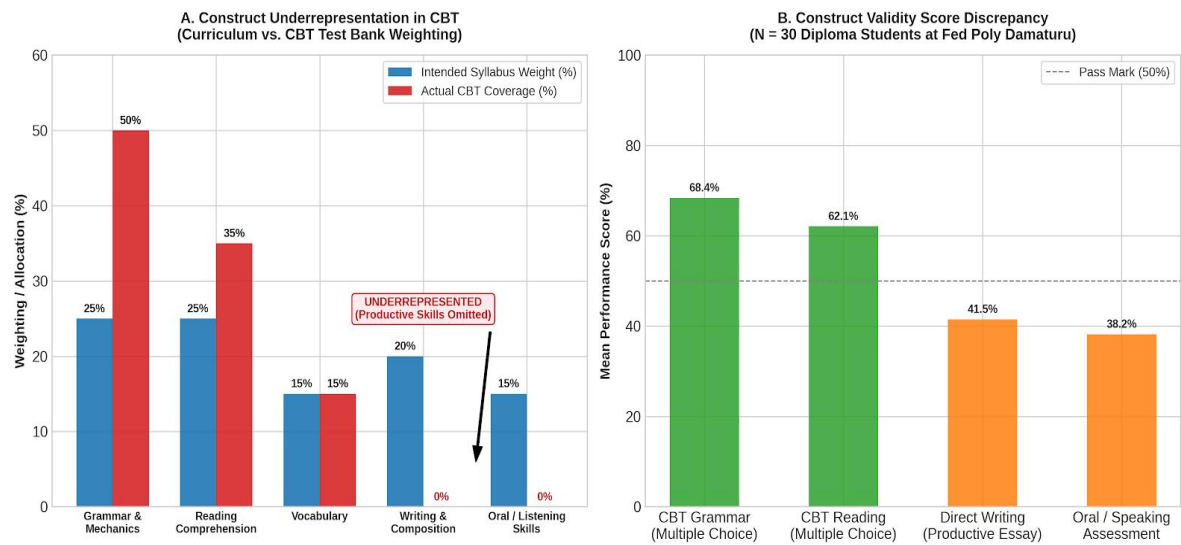


Table (A) Evaluating Construct Underrepresentation

Table (B) Evaluating Construct Validity

- Major findings based on the above Table A, Construct underrepresentation happens because 40% and above of the approved syllabus is totally misplaced from the CBT arrangement due to interface boundaries and dependence on computerized multiple-choice questions (MCQs). The impact is obvious by reallocation of deleted domains which expands the weighting of lower-order skill acknowledgment tasks.
- The other chart indicates an important performance crevasse between CBT acknowledgement tasks and direct productive tasks. The Low Convergent Validity on the CBT multiple-choice grammar items do not relate with students' capability in productive writing or oral communication. The CBT score reproduces test-wiseness and acknowledgement retention rather than genuine English language skill.

Analyzing this exam corpus from a psychometric viewpoint discloses critical intuitions concerning construct validity and construct underrepresentation.

1. Analysis of Construct Underrepresentation

Construct underrepresentation happens when an evaluation instrument leaves out the essential dimensions or other sub-skills of the target curriculum, which makes the test overly narrow. Some of the questions mainly test hypothetical, operational, and terminology-inclined knowledge.

Similarly, some of the tests evaluates declarative knowledge rather than procedural language skills by the use of the term critically analyze and synthesize. Generally, the courses under Use of English language target to construct functional reading comprehension, so, the test heavily focus on grammatical labels and some slight literary terminologies which genuinely underrepresents actual language competence.

The study also established a sort of optionality that undermines domain sampling. The use of forced-choice prompts such as question one is compulsory, answer question one and any other three or you either attempt A or B... permits the students to avoid or bypass the complete semantic or structural domains. For instance, a candidate may choose sentence according to its structure and completely avoid showing knowledge of Sentence according to purpose. A student may avoid testing on types of clauses by choosing types of phrases.

- **Validity Impact:** Non-compulsory questions may create a non-comparable test forms across distinct students and systematically permit students to skip the key curriculum areas, which may directly cause construct underrepresentation.
- **Structural Exclusion of Receptive Competencies:** most of the corpus exclude reading comprehension passages, listening/reading, summary writing tasks and other skill assessment. This complete absence isolates major pillars of language skills.

Table 1: Construct Assessment

Cognitive Target	Psychometric Evaluation	Primary Validity Issue
Clauses vs. Phrases	Tests operational arrangement via voluntary selections.	Construct Underrepresentation: Selection allows avoiding entire operational domains.

Cognitive Target	Psychometric Evaluation	Primary Validity Issue
Academic Pre-writing Skills	Real-world, technical implementation of constricting topics and outlining.	High Construct Validity: Direct operative measure of academic writing competence.
Vocabulary & Register (Technical Jargon)	Relates definition and context-inclined occupational jargon.	Construct-Irrelevant Variance: Score may be contingent on student's precise academic domain.
Drama Terminology & Types	Rote recollection of literary descriptions and list of 10 terms.	Construct Underrepresentation: Recompenses memory recollection over literary/critical investigation.

Entire Absence of Direct Language Performance:

Despite being a predominantly communication examination, most of the papers require zero real communicative performance. Testing Knowledge on some common terms and formats render the assessment strictly to underrepresents the core construct of professional communication. This substantial dependence on enumeration seeks to sample only lower-order cognitive stages while omitting higher-order critical analysis, application, editing and discourse synthesis.

2. Analysis of Construct Validity

Construct validity scrutinizes whether the test items produce answers that sincerely replicate the target fundamental trait. Some questions exhibit high construct validity for academic writing mechanics. Instead of just asking students to define an outline or thesis statement. Theoretically, it forces candidates to carry out the actual procedure of academic essay planning. It tests deployed, high-order cognitive proficiency. Other forms of the question may ask students to discuss jargons under different disciplines, in a heterogeneous student population which may include Business, Engineering, Sciences, Health, Management and other related field, may result to scoring inconsistency and makes vulnerable to prompt unpredictability. Students in fields with compressed, simply recognizable jargon may have an unfair advantage over students in universal fields, presenting construct-irrelevant variance. Construct validity assess whether test scores highlight the true capability in the directed construct while avoiding the distortion of the measurement errors. Because repeatedly sampling of the basic textbook definitions of the communication model, the assessment bank may lack the requisite to differentiate between a student who has only memorized a common diagram in a textbook and the one who attempt to apply communication tenets to solve practical workplace problems (Febriyana & Harjanto, 2023). Additionally, the fundamental annexation of optional prompts such as the clauses, phrases and other grammatical markers allow the students to strategically avoid other sub-domain of the syllabus, and later compromise the general domain coverage (Bartolome & Navalta, 2025).

Conversely, the assessment highlights discrete disparity in item excellence, which ranges from high construct validity to the basic form of construct irrelevant variance. This task serve as a direct performance of the common form of procedural composition mechanics which aligns closely with task-inclined evaluation tenets and argument-inclined authentication models (Xie, 2024). In distinction, the item which evaluates register and vocabulary through professional jargon may seek to introduce peripheral noise by querying students to explain technical jargons within their particular area of scholarship. In a diverse examination cluster, students from fields with dense highly effective terminologies may gain an unfair merit over their peers in wider disciplines, thereby assessing field-restricted background knowledge as against pure language skills (Messick, 1989; Pengelley et al., 2025).

Eventually, the psychometric examination offered in the table underlines the obligation of reorganization of high-stakes examination papers to guarantee defensible test score clarifications. When evaluations is dependent upon rote recollection and optional item structures, high test scores may fail to assure functional language capability in academic and occupational domains (Chapelle & Lee, 2021; Leung, 2025). To improve construct validity, potential test development in the polytechnic sector must minimize optionality, replace isolated metalinguistic prompts and restrict low-level listing tasks with combined performance tasks that assess receptive comprehension, genuine written production and functional grammar (Alshurafat et al., 2024; Ravand, 2025). Supporting examination stimuli with higher-order reasoning depth guarantees that test results serve as precise, all-inclusive indicators of student competence Agtarap et al., 2024).

Conclusion

The findings demonstrate that most measurement tools failed to show high coherence and assessing examination instruments via psychometric models such as construct validity discloses that depending heavily on lower-order metalinguistic recollection, isolated terminology listing and other basic structures may seriously affect the test score integrity whereas the isolated items which include procedural academic pre-writing tasks may show powerful genuine validity, the broad lack of direct productive composition, integrated language tasks and critical reading may lead to systematic

construct underrepresentation and construct-irrelevant variance. So, finally, aligning examination elements with genuine language use guarantees that potential assessments may deliver reliable, valid and psychometrically defensible interpretations concerning candidates' true academic and professional competences.

Ethical Considerations

The study was carried out in line with the current institution's regulations, laws and principles of NBTE. The data for this study were anonymized because it was obtained from the institution's examination database.

References

- Abate, T., & Mishore, E. (2024). Alignment analysis between teacher-made tests with the learning objectives in a selected school of central regional state of Ethiopia. *Heliyon*, 10(11)
- Adebola, O. O. (2022). Pre-Service Teachers Experiences during Teaching Practice in South Africa: Challenges and Solutions. *Journal of Education, Teaching and Learning*, 7(1):66-72 <https://dx.doi.org/10.26737/jetl.v7i1.3146>
- Agarap, M., Varier, D., & Constantinou, E. (2024). Evaluating teacher-made test items and assessment preparation in higher education. *International Journal of Research and Innovation in Social Science*, 8(4), 1531–1547.
- Al-Mekhlafi, M. A. A., & Naji, I. M. (2013). Supervisors' practices in improving student-teachers' grammar teaching strategies during practicum: A comparison of student teachers' and supervisors' views. *Wudpecker Journal of Educational Research*, 2(1), 6–17. <https://www.researchgate.net/publication/287331834>
- Alshurafat, H., Al Shbail, M. O., & Masadeh, W. M. (2024). Cognitive level alignment of accounting examination questions using Bloom's Revised Taxonomy: A content analysis approach. *Journal of Accounting Education*, 68, Article 100912. <https://doi.org/10.1016/j.jaccedu.2024.100912>
- Anderson, J. (2018). The role of initial teacher training courses in the professional development of experienced non-native speaker English language teachers. *English Language Teacher Education and Development*, 21.a. http://www.elted.net/uploads/7/3/1/6/7316005/-anderson_2018.pdf
- Ariza-Quñones, K. L., Hernández-Polo, L. D., Lesmes-Lesmes, K. J., & Molina-Ramírez, E. L. (2022). Preservice EFL teachers' experiences in their first teaching practicum: A collaborative autoethnography. *Profile Issues in Teachers' Professional Development*, 24(1), 193–207. <https://revistas.unal.edu.co/index.php/profile/article/view/91259>
- Bartolome, D. J. P., & Navalta, J. D. (2025). Misalignment in cognitive levels: A taxonomic analysis of Tables of Specification and mathematics examinations. *Journal of Education, Learning, and Management*, 2(2), 194–207. <https://doi.org/10.69739/jelm.v2i2.1025>
- Benton, S.L & Ryalls, R. K. (2016). Challenging misconceptions about students' ratings of instruction [IDEA Paper 55. https://www.ideaedu.org/idea_papers/challenging-misconceptions-about-student-ratings-of-instruction/
- Benton, S.L & Young, S. (2018). Best practices in the evaluation of teaching. IDEA Paper #69. <https://files.eric.ed.gov/fulltext/ED588352.pdf>
- Botaccio, L. A., Ortega, J. L. G., Rincón, A. N., & Fuentes, A. R. (2020). Evaluation for teachers and students in higher education. *Sustainability*, 12(10), 4078. <https://doi.org/10.3390/su12104078>
- Bowen, G. A. (2023). Document analysis as a qualitative research method. *Qualitative Research Journal*, 23(1), 14–28. <https://doi.org/10.3316/qj0902027>
- Chapelle, C. A., & Lee, Y. W. (2021). Argument-based validity in language testing: Foundations and applications. *Language Testing*, 38(1), 45–68. <https://doi.org/10.1177/0265532220925487>
- Febriyana, A., & Harjanto, I. (2023). Cognitive level analysis of English semester examination questions based on Revised Bloom's Taxonomy. *Journal of English Language Teaching and Linguistics*, 8(2), 145–160.
- Gupta, S., Grover, S., Menon, V., Indu, P. V., Vidhukumar, K., & Chacko, D. (2025). Item generation and establishing face and content validity of a rating scale: A primer. *Indian Journal of Psychiatry*, 67(8), 816–822. https://doi.org/10.4103/indianjpsychiatry_750_25
- Khoy, M. A. (2025). Evaluating the cognitive complexity of higher education language assessment tasks: A content analysis using Bloom's Revised Taxonomy. *System*, 128, Article 103510.
- Leung, C. (2025). Test validation: Beyond arguments in communicative language assessment. *Language Teaching Research Quarterly*, 50(11), 1–18.
- Menon, V., Grover, S., Gupta, S., Indu, P. V., Chacko, D., & Vidhukumar, K. (2025). A primer on reliability testing of a rating scale. *Indian Journal of Psychiatry*, 67(7), 725–729. https://doi.org/10.4103/indianjpsychiatry_584_25
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). Macmillan Publishing Company.
- OECD. (2023). *Digital assessment: OECD Digital Education Outlook 2023*. OECD Publishing. <https://doi.org/10.1787/c74f03de-en>

- Pengelly, R., Yeboah, A., & Weirich, S. (2025). Computer-based testing in higher education: A phenomenological investigation into undergraduate students' perspectives through the Technology Acceptance Model. *BMC Education*, 25(1), 102–118.
- Ravand, H. (2025). Measuring cognitive levels in high-stakes testing: Item characteristic curve and CDM analysis of university entrance examinations. *Frontiers in Education*, 10, Article 1644811. <https://doi.org/10.3389/educ.2025.1644811>
- Schneider, I. (2017). Can we trust measures of political trust? Assessing measurement equivalence in diverse regime types. *Social Indicators Research*, 133(3), 963–984.
- Shou, Y., Sellbom, M., & Chen, H.-F. (2022). Fundamentals of measurement in clinical psychology. In G. J. G. Asmundson (Ed.), *Comprehensive Clinical Psychology* (2nd Edition) (pp. 13–35). Elsevier. <https://doi.org/10.1016/B978-0-12-818697-8.00110-2>
- Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45, 83–117. <https://doi.org/10.1023/a:1006985528729>
- Sportel, B. E., Aardema, H., Boonstra, N., Arends, J., Rudd, B., Metz, M. J., Castelein, S., & Pijnenborg, G. H. M. (2023). Measuring recovery in participants with a schizophrenia spectrum disorder: Validation of the Individual Recovery Outcomes Counter (IROC). *BMC Psychiatry*, 23(1), 296–296. <https://doi.org/10.1186/s12888-023-04763-3>
- Tsagari, D., & Bohn, H. (2025). Preface to "Recent developments in language testing and assessment". *Languages*, 10(2), Article 22. <https://doi.org/10.3390/languages10020022>
- Xie, Q. (2024). Construct representation and predictive validity of integrated writing tasks: A study on automated language assessments. *Assessing Writing*, 61(4), Article 100846. <https://doi.org/10.1016/j.asw.2024.100846>