Cakrawala Pendidikan
Jurnal Ilmiah Pendidikan
Vol. 45 No. 2, June 2026, pp. 458-466
https://journal.uny.ac.id/index.php/cp/issue/view/3110
DOI: https://doi.org/10.21831/cp.v45i2.95780

Developing a proficiency test of Academic English in the Indonesian higher education context

Didi Sukyadi*, Fuad Abdul Hamied, Ika Lestari Damayanti, Ari Arifin Danuwijaya, Lukman Hakim, Firly Asyifa
Universitas Pendidikan Indonesia, Indonesia
*Corresponding Author: sukyadi.d.upi@gmail.com

ABSTRACT

Standardized tests remain crucial for students whose first language is not English in today's globalized world, despite ongoing debates about their suitability. The purpose of this research is to develop an academic English proficiency test for higher education in Indonesia. The data were derived from a needs analysis conducted through focus group discussions (FGDs) and a pre-survey for the existing academic English proficiency test, followed by a test design process: creating a test blueprint and developing question items, and a second survey following a trial test with 538 test takers. The findings indicate that adjustments were required to the test format and duration, a basis for developing a new test format. Additionally, students reported a preference for computer-based tests. During the test design process, the research team revised the test, retained the test objectives, and modified the test item types. Lastly, post-test feedback and item analysis indicated an imbalance in item difficulty in the Structure and Written Expression section, while the Academic Listening and Academic Reading sections demonstrated relatively balanced difficulty levels. Overall, the findings of this research indicate a desire to revisit and refine the existing test to better meet students' needs while maintaining its objectives.

Keywords: Academic English, English language testing, higher education, needs analysis, test development

Article history

Received:	*Revised:*	*Accepted:*	*Published:*
09 February 2026	*14 March 2026*	*03 May 2026*	*29 May 2026*

Citation (APA Style): Sukyadi, D., Hamied, F. A., Damayanti, I. L., Danuwijaya, A. A., Hakim, L., Asyifa, F., (2026). Developing a proficiency test of Academic English in the Indonesian higher education context. *Cakrawala Pendidikan: Jurnal Ilmiah Pendidikan, 45*(2), pp. 458-466. DOI: https://doi.org/10.21831/cp.v45i2.95780

INTRODUCTION

The use of the English language has grown significantly across many spheres of life, particularly in today's globalized world (Sadeghpour & Sharifian, 2019). In this context, English functions as an international lingua franca, enabling communication among people from different countries and linguistic backgrounds (Cappuzzo, 2024; Pennycook & Candlin, 2017). English has evolved into a contact language used by speakers who do not share a common first language, leading to the emergence of new varieties of English (Ojha, 2022). As English is used across diverse contexts, the demand for effective and appropriate English-language practices has increased, requiring textbook writers, teachers, and test developers to pay closer attention to the needs of non-native English speakers.

This development aligns with growing critiques of approaches and assessment tools guided by a monolingual ideology of English, which are increasingly questioned in today's multilingual world. Ojha (2022) notes that there is a growing need to design English proficiency tests that can accommodate learners' diverse linguistic and cultural backgrounds. However, this does not imply that English learners must master all existing varieties of English; rather, proficiency in academic contexts requires the ability to comprehend and use English for lectures, academic texts, and written tasks. Rather, proficiency in contemporary contexts involves not only grammatical accuracy but also the ability to negotiate meaning with interlocutors from diverse linguistic,

cultural, and social backgrounds who use different varieties of English to communicate effectively (Canagarajah, 2006).

Even though continued debates about the value of the English proficiency test persist, it remains essential for students who are not native English speakers, particularly in higher education contexts where English proficiency is closely linked to academic participation. The fact that standardized English proficiency tests such as TOEFL, IELTS, PTE, and Cambridge ESOL are mandatory requirements for college admission for multilingual students in English-dominant countries such as Australia, the United Kingdom, and the United States indicates the importance placed on these tests by higher education institutions (Zhang-Wu & Brisk, 2021). Across Asia, including Taiwan (Hsieh, 2017), Korea (Choi, 2008), Hong Kong (Qian, 2008), and Vietnam (Nhan, 2013), English proficiency is widely recognized as an important component of academic readiness and students' ability to engage in English-medium academic practices. Furthermore, in Indonesia, the Ministry of Education, Culture, Research, and Technology issued Regulation No. 3 of 2020 on the National Standards for Higher Education, highlighting English proficiency as an indicator of graduate readiness. Thus, as argued by Hamid et al. (2019), standardized English proficiency tests function as gatekeepers of higher education.

One way that can be done to improve such tests is by using digital platforms. In the Indonesian ELT context, digital platforms have been proven to support computer-based assessment. Their advantages include being more objective, fair, accurate, valid, and reliable, while also reducing opportunities for cheating (Sumardi & Muamaroh, 2020). This indicates that academic English proficiency tests need to be redesigned to be more flexible and technology based. Similar challenges are also evident in the development of ESP courses. The main challenges include differences in students' English proficiency, the relevance of materials, and alignment with academic or professional needs in designing effective English programs in higher education (Huang et al., 2024). To address these challenges, a recent study in the Indonesian ESP context found that technology-based English learning can improve students' readiness, motivation, self-regulated learning, and confidence in professional communication tasks (Winantaka et al., 2025). These findings further reinforce that academic English proficiency tests should be designed based on students' actual academic and communicative needs.

However, this recognition underscores the need to reconsider how academic English proficiency is conceptualized and assessed. In this regard, standardized English proficiency tests need to be revisited and adapted to better reflect the realities of English use in multilingual and multicultural contexts. Such reconsideration requires a shift in assessment practices that can respond to changing purposes, users, and contexts of English use (Ojha, 2022), as well as a critical examination of established testing practices and underlying assumptions (Hu, 2012).

Additionally, extensive discussions on reconceptualizing English language assessment in multilingual contexts, as well as empirical studies systematically developing, validating, and examining the reliability of standardized academic English proficiency tests, remain limited, particularly in higher education settings. Wang and Li (2020) note that research on multilingual assessment has largely focused on conceptual discussions, ideological shifts, and classroom-based practices, while concrete examples of test development, validation, and implementation remain scarce. Similarly, although Schissel et al. (2018) provide empirical evidence that assessment designs incorporating multilingual resources enable learners to demonstrate higher-order thinking and more complex language abilities, their work is situated primarily within classroom-based, task-based assessments. As such, there remains a gap in the development of standardized academic English proficiency tests that can function reliably and validly in high-stakes higher education contexts.

To address this gap, it is necessary to revisit the principles of English for Academic Purposes (EAP) testing, which provide a relevant framework for aligning academic English assessment with students' actual academic needs. From an EAP perspective, the quality of an academic English proficiency test is determined not only by its technical soundness but also by its relevance to students' academic needs. EAP test should have a clear target, namely, to determine how well students can use English in their academic subject areas and should reflect real-life academic contexts in which the language is used (Damayanti et al., 2024). Central to this

principle is the needs analysis, as Hughes (2003) emphasizes that EAP tests must be grounded in a systematic investigation of learners' academic and linguistic needs. EAP tests should simulate the academic tasks students are expected to perform in their respective disciplines (Song & Zhou, 2022). In addition, a sound EAP test must demonstrate validity and reliability and be capable of measuring communicative competence, including grammatical, discourse, sociolinguistic, strategic, and cultural competence (Chemir & Kitila, 2022). To address this gap, this research aims to improve the reliability of an academic English proficiency test for higher education students in Indonesia. This purpose is addressed through the following research questions: (1) What needs do students identify regarding the format, duration, and delivery mode of the existing academic English proficiency test, and how can these needs inform the redesign of the test? (2) To what extent does the revised academic English proficiency test demonstrate preliminary validity and reliability based on pilot testing and test-taker feedback?

METHOD

This study used a Research and Development (R&D) design to develop and produce a valid and reliable website-based PTESOL English language proficiency test instrument. The approach used is Plomp's (Faisal et al., 2024; Iriani et al., 2023) Design-Based Research (DBR) model, an iterative, empirically grounded model. This design is particularly suitable for developing complex, contextually grounded assessment instruments because it allows researchers to continuously refine the product based on data and feedback obtained from each stage of development.

This research was designed in four main phases in accordance with the DBR model: (1) Needs and Context Analysis, (2) Prototype Design and Development, (3) Pilot Implementation and Evaluation, and (4) Validation and Refinement. At the time of writing, the study has completed the first three phases and is currently preparing for the fourth phase. This design ensures that the instruments developed not only meet user needs but also meet academic and professional quality standards.

The needs analysis phase involved undergraduate and postgraduate students from non-English language majors enrolled in higher education institutions who were required to take an academic English proficiency test as part of their academic requirements. Qualitative data were obtained through focus group discussions (FGDs), while quantitative data were gathered through a pre-survey administered to students who had previously taken the existing academic English proficiency test.

For the pilot phase, the revised test was administered to 538 test takers drawn from higher education contexts in Indonesia. These participants represented the target population of the proficiency test. Following the trial test, a post-test survey was conducted to collect test-takers' perceptions of the revised test.

This study employed four main research instruments to collect both quantitative and qualitative data. First, a needs analysis questionnaire was designed to gather information on respondents' demographic characteristics, the context and frequency of English use in academic activities, perceptions of the relevance and appropriateness of the existing English proficiency test, technical aspects, and balance among test sections, and expectations regarding the characteristics of an ideal academic English proficiency test. The questionnaire was developed based on the needs analysis framework which differentiates between learners' target needs, including necessities, lacks, and wants, and their learning needs.

Second, FGD guidelines were developed to obtain in-depth qualitative insights into students' experiences with the existing standardized academic English proficiency test and their expectations of an ideal test instrument. Three FGDs were conducted at different stages of the test development process: (1) the first FGD focused on identifying key considerations for test instrument development, (2) the second FGD was conducted to review and refine the test blueprint draft, and (3) the third FGD aimed to evaluate the test prototype and gather feedback for further revision.

Third, an expert validation sheet was employed and completed by two language assessment experts to evaluate the test instrument's quality. The validation focused on several aspects, including the relevance of test content to the intended construct (content validity), the alignment of test items with the test specifications (construct alignment), the clarity of language and instructions, the technical quality of distractors, and the appropriateness of the difficulty level for the target population.

Finally, a test participant feedback questionnaire was administered following the pilot test to collect participants' perceptions of the test, particularly regarding the clarity of instructions, level of difficulty, relevance of content to academic needs, and adequacy of the allotted time.

In line with the R&D approach, this study was conducted through multiple phases, including needs analysis, test design, and pilot implementation.

The first phase aimed to identify non–English-major students' academic English needs, patterns of language use in academic contexts, and perceptions of the existing English proficiency test. Data were collected through an initial FGD with students from various study programs to explore their experiences with academic English, the challenges they faced, and their expectations for an English proficiency test. Key issues identified included test duration and accessibility.

A needs analysis survey was subsequently administered to examine technical constraints, content relevance, and balance among test sections in the existing test. To ensure construct relevance, the study also included benchmarking with institutions experienced in administering web-based English proficiency tests, analysis of students' English-mediated academic tasks, and a review of international and local test frameworks (e.g., TOEFL iBT, IELTS, previous PTESOL, and TOEP).

Guided by the findings of the needs analysis and the Common European Framework for Reference for Languages (CEFR) framework, a test blueprint was developed comprising three constructs, Academic Listening, Structure and Written Expression (SWE), and Academic Reading, with 60 items and a test duration of 65 minutes. Test development involved defining the construct, specifying the blueprint, writing items with varied difficulty levels, and expert validation by language assessment specialists focused on content relevance, construct alignment, and technical quality. A web-based test prototype was subsequently developed, incorporating automated item delivery, randomization, scoring, and participant data management. However, the pilot test was administered in paper-based form to examine item quality independently of technological factors.

A paper-based pilot test was conducted to evaluate item quality independent of technological factors. Quantitative data were analyzed using Classical Test Theory (CTT) to examine item difficulty, discrimination, and test reliability using the Kuder–Richardson Formula 20 (KR-20). Qualitative data from participant feedback questionnaires and post-pilot FGDs were analyzed thematically to identify issues related to clarity, difficulty level, and content relevance.

The final phase will involve large-scale administration of the revised test to conduct construct validation using Confirmatory Factor Analysis (CFA), finalize test refinement, establish cut-off scores through standard-setting procedures (e.g., the Angoff method), and prepare the instrument for full-scale institutional implementation.

Qualitative data obtained from FGDs, open-ended survey responses, and test participant feedback were analyzed using thematic analysis. The analysis involved data familiarization through repeated reading and full transcription of FGD recordings; initial coding to identify meaning units relevant to the research objectives; theme identification and clustering based on emerging patterns; theme review and labeling to ensure coherence and relevance; and reporting of findings with illustrative excerpts to provide contextualized evidence.

Quantitative data from the pilot test were analyzed using descriptive statistics and CTT procedures. Item difficulty was calculated using the percentage-correct formula (p = number of correct responses / total responses) and classified as difficult ($p < .40$), moderate ($.40 \leq p \leq .70$), or easy ($p > .70$). Item discrimination was examined using the upper–lower group comparison method, calculated as the difference between the proportions of correct responses in the upper and lower 27% groups, and interpreted as very good ($D \geq .40$), good (.30–.39), marginal (.20–.29), or poor ($D < .20$).

Test reliability was estimated using the KR-20 to assess internal consistency for dichotomously scored items. A minimum reliability coefficient of .70 was considered acceptable for low-stakes testing, while a threshold of .85 was applied for high-stakes contexts such as English proficiency tests. Content validity was evaluated using the Content Validity Index (CVI) based on expert judgments of item relevance on a four-point scale, with a minimum acceptable CVI of .80.

FINDINGS AND DISCUSSION

Findings

Students' needs related to test format, duration, and delivery mode

The needs analysis survey involved students with a mean age of 26.5 years and an average study duration of 6.6 semesters, representing undergraduate (67.5%), master's (22.8%), and doctoral (9.7%) levels from various non–English-related disciplines, including education and management. This profile indicates that the respondents had sufficient academic experience to articulate their needs in academic English.

The analysis identified three main contexts of academic English use. First, students reported a strong need for academic listening skills, particularly for understanding lectures, instructional videos, seminars, and conference presentations delivered in English. Second, Academic Reading emerged as a core requirement, as students frequently engaged with authentic academic texts such as journal articles and textbooks to complete coursework and produce theses or dissertations. Third, students emphasized the importance of grammatical accuracy and appropriate lexical choice in academic writing, especially for abstracts, research reports, and proposals. These findings reflect the cognitively demanding and decontextualized nature of academic language use and support a genre-based view of academic literacy.

Regarding the existing English proficiency test, most respondents (78%) considered the assessed skills relevant; however, several limitations were identified. The two-hour test duration was considered excessively long and cognitively demanding, with 65% of respondents preferring a more time-efficient format. Accessibility was another major concern, as limited test locations and computer availability led to long waiting times; consequently, 82% of respondents expressed a preference for more flexible access options. Technical issues, particularly related to audio quality (23%) and occasional system instability, were also reported. In addition, 43% of respondents indicated a need for diagnostic feedback to help identify areas for improvement. Finally, students expressed a clear preference for more authentic academic tasks, such as critical reading of journal articles, that better reflect real academic demands.

Overall, these findings provide empirical justification for developing an English proficiency test that is more efficient, accessible, technically reliable, and aligned with authentic academic language use, consistent with user-centered and practicality-oriented assessment principles.

Item analysis

Item analysis was conducted using Classical Test Theory (CTT) based on data from the second pilot administration. Overall, the results revealed an uneven distribution of item difficulty across test sections, with a general tendency toward higher difficulty levels.

In the Academic Listening section, item difficulty was relatively well balanced, with a majority of items falling within the moderate range (60%), alongside a smaller proportion of difficult items (35%) and one easy item (5%). Moderately difficult items primarily targeted main ideas and explicit details in long conversations, while more difficult items required inferential comprehension in short talks. Although the proportion of difficult items slightly exceeded the recommended range, the section overall demonstrated an acceptable balance for an exit-level test.

The SWE section exhibited the most pronounced imbalance, with no easy items, a limited number of moderate items (25%), and a predominance of difficult items (75%). Items that were extremely difficult were typically associated with complex grammatical structures or highly

specialized academic vocabulary. This pattern suggests a potential floor effect, which may reduce measurement precision and undermine validity, indicating a clear need for revision.

The Academic Reading section showed a more favorable distribution, with items spread across easy (15%), moderate (50%), and difficult (35%) categories. Easy items generally assessed explicit information, moderate items focused on inter-sentential relationships and interpretation, and difficult items required critical evaluation or synthesis across texts. This distribution aligns reasonably well with psychometric recommendations for effective discrimination.

Item discrimination was examined using the upper–lower 27% group comparison method. Mean discrimination indices were satisfactory across sections: Academic Listening (D = 0.38), Structure and Written Expression (D = 0.31), and Academic Reading (D = 0.42). Most items demonstrated acceptable to strong discrimination (D ≥ 0.30), particularly in the Academic Reading section. However, several items in the Structure and Written Expression section showed marginal discrimination, further supporting the need for targeted revision.

Test reliability, estimated using the KR-20 coefficient, was strong. 0.76 for Academic Listening, 0.72 for Structure and Written Expression, 0.81 for Academic Reading, and 0.87 for the overall test, and 0.87 for the overall test. The overall reliability exceeded the recommended threshold for high-stakes testing, while subtest reliabilities met acceptable standards, with the lowest reliability observed in the Structure and Written Expression section.

In summary, the item analysis demonstrates that the revised test possesses generally satisfactory psychometric properties, particularly in the Academic Listening and Academic Reading sections. Nonetheless, a systematic revision of the Structure and Written Expression section is necessary to improve the balance of difficulty, discrimination, and overall measurement precision before large-scale implementation.

Discussion

The item analysis results indicate that the prototype academic English proficiency test demonstrates an adequate overall range of difficulty levels; however, the distribution of item difficulty is not yet optimal, particularly in the Structure and Written Expression (SWE) section. While the Academic Listening and Academic Reading sections show relatively balanced distributions across easy, medium, and difficult items, the SWE section shows a strong skew toward difficult items, with no items at the easy level and a large majority classified as difficult. Such an imbalance has important implications for the test's psychometric quality, fairness, and educational impact.

From a psychometric perspective, an excessive concentration of difficult items threatens construct validity because the test does not measure the target construct with equal precision across the full range of examinee ability. A valid proficiency test should capture meaningful variation among lower-, middle- and higher-ability learners; however, when most items are very difficult, the test primarily provides information about higher-ability examinees while yielding limited diagnostic value for those at intermediate or lower levels (Al-Zboon et al., 2021; Tan, 2024). As many university-level test takers typically fall within the intermediate to upper-intermediate range, the SWE section in its current form may not adequately represent their grammatical competence. Moreover, a highly skewed difficulty distribution increases the likelihood of a floor effect, in which a large proportion of test takers obtain very low scores. This restriction of score variance can reduce test reliability, as the instrument’s ability to consistently differentiate among examinees diminishes when scores cluster at the lower end of the scale. Although the overall reliability of the test meets accepted standards for high-stakes assessment, the SWE section's lower reliability suggests that its contribution to measurement precision could be improved through targeted revision.

Item difficulty is also closely linked to item discrimination. Items with moderate difficulty levels tend to yield the highest discrimination indices, whereas very difficult items often fail to effectively distinguish between higher- and lower-ability test takers because even strong examinees struggle to answer them correctly. While the average discrimination of the SWE section remains acceptable, the presence of marginally functioning items indicates that a more

balanced difficulty profile would likely enhance the section's discriminative power and overall measurement quality.

Beyond psychometric concerns, the imbalance in item difficulty raises important issues related to fairness and test acceptability. Tests perceived as excessively difficult may increase test anxiety and undermine examinees' confidence, which, in turn, can negatively affect performance not only in the difficult section itself but also in subsequent sections due to carryover effects. When examinees feel that a test does not reflect attainable or relevant language demands, they may question its legitimacy and usefulness, particularly when the test is used for high-stakes decisions such as graduation or academic progression. Reduced stakeholder acceptance can ultimately weaken trust in institutional assessment practices and policies.

The imbalance in difficulty also has potential implications for washback. An academic English proficiency test that places disproportionate emphasis on highly complex grammatical structures may encourage teaching and learning practices that prioritize test-taking strategies and mastery of rare or overly technical forms at the expense of developing broader academic language competence. Such negative washback has been widely documented in the language testing literature, particularly when assessment practices are misaligned with instructional goals. Students who repeatedly fail a grammar-focused section despite being able to function adequately in academic contexts may become demotivated and disengaged, perceiving English as an arbitrary barrier rather than a tool for academic success. Similarly, instructors may feel compelled to "teach to the test" by devoting excessive instructional time to complex grammatical drills, potentially distorting curricula and diminishing attention to equally important skills such as academic reading, writing, and critical engagement with texts.

Addressing these issues requires a systematic revision of the SWE section to achieve a more proportional distribution of item difficulty. Increasing the number of easy and medium-level items would improve construct representation, enhance reliability, and reduce the risk of floor effects, while still retaining sufficient difficult items to challenge higher-ability examinees. Revision strategies may include simplifying overly complex grammatical constructions, replacing rarely used or highly technical forms with common structures in academic discourse, and providing clearer contextualization to reduce unnecessary cognitive load. Research suggests that contextualized grammar items offer a more valid representation of grammatical ability in real language use than decontextualized items. In addition, developing new items specifically targeting easy and medium difficulty levels, followed by expert validation and pilot testing, would help ensure that revisions enhance rather than compromise psychometric quality.

The iterative nature of the design-based research approach adopted in this study provides a strong methodological foundation for such revisions. By treating pilot testing and item analysis as integral components of an ongoing development cycle, this approach enables data-driven refinement rather than one-time test construction. Stakeholder input, expert judgment, and empirical evidence collectively inform revision decisions, ensuring that the test evolves toward greater validity, reliability, and practical relevance. This stands in contrast to linear test development models that offer limited opportunities for systematic improvement once a test is operationalized.

A further strength of the test lies in its alignment with the Common European Framework of Reference for Languages (CEFR) while remaining responsive to context-specific academic language needs. CEFR provides an internationally recognized framework that supports conceptual clarity, score interpretability, and comparability across contexts. By grounding test design in CEFR descriptors, particularly those relevant to academic study, the test gains global legitimacy without sacrificing relevance. At the same time, the use of needs analysis ensures that the language abilities assessed reflect authentic academic tasks, such as understanding lectures, engaging with academic texts, and producing grammatically accurate academic writing. This balance between international standardization and contextual relevance aligns with recommendations in the language assessment literature that advocate an adaptive rather than rigid application of global frameworks.

Overall, the findings underscore the importance of balanced item difficulty in academic English proficiency testing, not only for psychometric robustness but also for fairness, acceptance,

and positive educational impact. While the current prototype demonstrates strong potential, particularly in its Academic Listening and Academic Reading components, targeted revision of the SWE section is essential to ensure that the test measures grammatical competence accurately across ability levels and supports constructive washback. With continued iterative refinement, empirical validation, and alignment between assessment and academic language demands, the academic English proficiency test developed in this study has the potential to function as a valid, reliable, and pedagogically responsible assessment instrument in higher education contexts.

CONCLUSION

This study developed a web-based prototype of an English proficiency test for non–English major university students using a Design-Based Research approach. The findings indicate that the existing English proficiency test posed substantial challenges in terms of test duration, accessibility, and user experience, while students expressed a clear need for a more efficient, flexible, and academically authentic assessment. In response, the proposed English proficiency test was designed around CEFR-aligned academic constructs and demonstrated strong content validity and overall reliability, indicating its potential to measure academic English proficiency effectively while addressing key practical limitations.

Nevertheless, the study also revealed areas requiring further refinement. In particular, the imbalance in item difficulty, most notably in the Structure and Written Expression section, suggests the need for systematic item revision and revalidation. In addition, as the instrument has thus far been evaluated only through paper-based pilot testing, further empirical evidence is required to confirm its construct validity, scalability, and performance in a fully web-based testing environment. The absence of standard-setting procedures and evidence of predictive validity also constrains its immediate application for high-stakes decision-making.

Future research should therefore focus on iterative test revision, large-scale web-based implementation, and advanced validation studies, including factor analysis, standard setting, and washback analysis. With continued refinement and rigorous validation, the English proficiency test developed in this study has the potential to serve as a valid, reliable, and contextually relevant assessment tool following completion of the final DBR phase, including construct validation and final refinement, and to contribute meaningfully to language assessment practices in higher education.

REFERENCES

Al-Zboon, H. S., Alrekebat, A., & Abdelrahman, M. B. (2021). The effect of multiple-choice test items' difficulty degree on the reliability coefficient and the standard error of measurement depending on the Item Response Theory (IRT). *The International Journal of Higher Education, 10*, 22. DOI: https://doi.org/10.5430/ijhe.v10n6p22

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly, 3*(3), 229-242. DOI: https://doi.org/10.1207/s15434311laq0303_1

Cappuzzo, B. (2024). Plurilingualism, multilingualism, and lingua Franca English in today's globalised world. *International Journal of Linguistics Literature and Culture*. DOI: https://doi.org/10.19044/llc.v11no1a1

Chemir, S., & Kitila, T. (2022). English for academic purposes learners' needs analysis: Language difficulties encountered by university students in Ethiopia. *Celtic : A Journal of Culture, English Language Teaching, Literature and Linguistics*. DOI: https://doi.org/10.22219/celtic.v9i1.20646

Choi, I. -C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing, 25*(1), 39–62. DOI: https://doi.org/10.1177/0265532207083744

Damayanti, I. L., Derinalp, P., Asyifa, F., & Suryatama, K. (2024). Exploring English for academic purposes program: Needs analysis and impact evaluation. *Studies in English Language and Education*. DOI: https://doi.org/10.24815/siele.v11i3.38329

Faisal, A. H., Anshori, D. S., Sastromiharjo, A., & Mulyati, Y. (2024). Designing a data literacy-based speaking skills assessment instrument for high school students. *JISAE: Journal of Indonesian*

Student Assessment and Evaluation. DOI: https://doi.org/10.21009/jisae.v10i1.43780

Hamid, M. O., Hoang, N. T. H., & Kirkpatrick, A. (2019). Language tests, linguistic gatekeeping and global mobility. *Current Issues in Language Planning, 20*(3), 226–244. DOI: https://doi.org/10.1080/14664208.2018.1495371

Hsieh, C. N. (2017). The case of Taiwan: Perceptions of college students about the use of the TOEIC® tests as a condition of graduation. *ETS Research Report Series, 2017(1)*, 1-12. https://files.eric.ed.gov/fulltext/EJ1168727.pdf

Hu, G. W. (2012). *Assessing English as an international language*. In L. Alsagoff, S. L. McKay, G. W. Hu, & W. Renandya (Eds.), Principles and practices for teaching English as an international language (pp. 123-143). Routledge.

Huang, Z., Mustakim, S., & Ghazali, N. (2024). Analysing the challenges of developing English for Specific Purpose (ESP) courses for Sino-Foreign cooperative educational programs. *Cakrawala Pendidikan: Jurnal Ilmiah Pendidikan, 43*(3), 765-772. DOI: https://doi.org/10.21831/cp.v43i3.64927

Hughes, A. (2003). *Testing for language teachers* (2nd Ed.). Cambridge. DOI: https://doi.org/10.1017/CBO9780511732980

Iriani, T., Anisah, A., Luthfiana, Y., Maknun, J., & Dewi, N. I. K. (2023). Analytical rubric development design for objective test assessment. *Proceedings of the 4th Annual Conference of Engineering and Implementation on Vocational Education, ACEIVE 2022, 20 October 2022, Medan, North Sumatra, Indonesia*. DOI: https://doi.org/10.4108/eai.20-10-2022.2328882

Nhan, T. (2013). The TOEIC® test as an exit requirement in universities and colleges in Danang City, Vietnam: Challenges and impacts. *International Journal of Innovative Interdisciplinary Research, 2*(1), 33–50.

Ojha, L. P. (2022). World Englishes, monolingual bias, and standardized tests in a multilingual world: Ideologies, practices, and the missing link. *Journal of NELTA, 27*(1-2), 88-105. DOI: https://doi.org/10.3126/nelta.v27i1-2.53197

Pennycook, A., & Candlin, C. N. (2017). The cultural politics of English as an international language. *Routledge*.

Qian, D. D. (2008). English language assessment in Hong Kong: A survey of practices, developments and issues. *Language Testing, 25*(1), 85-110. DOI: https://doi.org/10.1177/0265532207083746

Sadeghpour, M., & Sharifian, F. (2019). World Englishes in English language teaching. *World Englishes, 38*(1-2), 245-258. DOI: https://doi.org/10.1111/weng.12372

Schissel, J. L., Leung, C., López-Gopar, M., & Davis, J. R. (2018). Multilingual learners in language assessment: Assessment design for linguistically diverse communities. *Language and Education*, 167-182. DOI: https://doi.org/10.1080/09500782.2018.1429463

Song, Y., & Zhou, J. (2022). Revising English language course curriculum among graduate students: An EAP needs analysis study. *SAGE Open, 12*. DOI: https://doi.org/10.1177/21582440221093040

Sumardi, S., & Muamaroh, M. (2020). Edmodo impacts: Mediating digital class and assessment in English language teaching. *Cakrawala Pendidikan*, *39*(2), 319-331. DOI: https://doi.org/10.21831/cp.v39i2.30065

Tan, T. K. (2024). Evaluating assessment via item response theory utilizing information function with R. *The Quantitative Methods for Psychology*. https://doi.org/10.20982/tqmp.20.1.p033

Wang, Y., & Li, S. (2020). Issues, challenges, and future directions for multilingual assessment. *Journal of Language Teaching and Research, 11*(6), 914-919. http://dx.doi.org/10.17507/jltr.1106.06

Winantaka, B., Efendi, A., & Putro, N. H. P. S. (2025). Exploring flipped classroom for business English: Preliminary insights from student voices and reflections. *Cakrawala Pendidikan: Jurnal Ilmiah Pendidikan, 44*(3), pp.730-739. DOI https://doi.org/10.21831/cp.v44i3.65140

Zhang-Wu, Q., & Brisk, M. E. (2021). "I must have taken a fake TOEFL!": Rethinking linguistically responsive instruction through the eyes of Chinese international freshmen. *TESOL Quarterly, 55*(4), 1136-1161. DOI: https://doi.org/10.1002/tesq.3077