p-ISSN: 2963-7856 | e-ISSN: 2961-9890
Available online at jerkin.org/index.php/jerkin
Jurnal Pengabdian Masyarakat dan Riset Pendidikan
Volume 4 No 2, Oktober-Desember 2025, pp 9109-9116

Linking Item Analysis to Teaching and Learning:What English Test Data Tell Us about Teachers' Competencies

Suharyadi[1], **Shirly Rizki Kusumaningrum**[2]*
[1,2]Pendidikan Profesi Guru, Universitas Negeri Malang, Jl. Semarang No 5 Malang
E-mail: suharyadi.fs@um.ac.id
*Corresponding Author
https://doi.org/10.31004/jerkin.v4i2.3116

ARTICLE INFO

Article history
Received: 10 October 2025
Revised: 15 October 2025 2025
Accepted: 29 October 2025

Kata Kunci
Analisis Butir Soal; Kompetensi Guru; Teori Tes Klasik; Penilaian Literasi; Pendidikan Bahasa Inggris

Keywords
Item Analysis; Teacher Competence; Classical Test Theory; Assessment Literacy; English Language Education

ABSTRACT

Studi ini mengkaji kualitas psikometrik dari tes uji coba yang diberikan kepada 283 guru bahasa Inggris yang sedang bertugas di Indonesia, menggunakan prosedur Teori Tes Klasik (CTT). Analisis difokuskan pada lima indeks: validitas item, diskriminasi, tingkat kesulitan, fungsi pengalih perhatian, dan pemanfaatan item. Hasil mengungkapkan beberapa kelemahan dalam desain tes. Hanya 46% item yang menunjukkan validitas yang dapat diterima, sementara 54% gagal menyelaraskan secara bermakna dengan konstruk keseluruhan kompetensi guru. Lebih dari separuh item menunjukkan daya diskriminasi yang buruk, membatasi kemampuan tes untuk membedakan guru yang berkinerja tinggi dan rendah. Tingkat kesulitan tidak merata, dengan 72% item dikategorikan sebagai mudah, sehingga membatasi pengukuran keterampilan berpikir tingkat tinggi. Analisis pengalih perhatian menunjukkan bahwa tiga perempat opsi tidak berfungsi, mengurangi nilai diagnostik. Akibatnya, hanya 9% item yang dapat dipertahankan tanpa revisi. Terlepas dari kekurangan ini, temuan ini memberikan wawasan yang berguna tentang kompetensi guru. Meskipun pengetahuan linguistik dan pedagogi dasar terwakili dengan baik, keterampilan profesional tingkat tinggi seperti inferensi, sintesis, dan pedagogi terapan masih kurang dinilai. Hasil ini menyoroti pentingnya memperkuat literasi penilaian dan mengintegrasikan analisis butir soal ke dalam pengembangan profesional guru.

This study examines the psychometric qualities of a tryout test administered to 283 in-service English teachers in Indonesia, using Classical Test Theory (CTT) procedures. The analysis focused on five indices: item validity, discrimination, difficulty, distractor functioning, and item utilization. Results revealed several weaknesses in the test design. Only 46% of items demonstrated acceptable validity, while 54% failed to align meaningfully with the overall construct of teacher competence. More than half of the items showed poor discrimination, limiting the test's ability to differentiate high- and low-performing teachers. Difficulty levels were skewed, with 72% of items categorized as easy, thereby restricting the measurement of higher-order thinking skills. Distractor analysis indicated that three-quarters of options were non-functioning, reducing diagnostic value. Consequently, only 9% of items could be retained without revision. Despite these shortcomings, the findings provide useful insights into teacher competencies. While basic linguistic and pedagogical knowledge was well represented, higher-order professional skills such as inference, synthesis, and applied pedagogy were under-assessed. These results highlight the importance of strengthening assessment literacy and integrating item analysis into teacher professional development.

How to Cite: Suharyadi, et al (2025). Linking Item Analysis to Teaching and Learning:What English Test Data Tell Us about Teachers' Competencies 4(2) 9109-9116 https://doi.org/10.31004/jerkin.v4i2.3116

INTRODUCTION

Assessment in education has often been perceived as a separate domain from classroom teaching, serving primarily as a way to measure achievement rather than to inform instruction (Bachman, 2022; Estaji, 2024; Varier et al., 2024). In language learning, particularly English as a Foreign Language (EFL), teachers routinely administer tests, yet the wealth of information embedded in students' responses is rarely explored beyond assigning grades. Item analysis—the process of evaluating test questions in terms of difficulty, discrimination, and distractor quality—offers a systematic way of looking at how students interact with assessment tasks (Rejeki et al., 2023). However, its potential is often underutilized.

Too often, item analysis remains confined to the realm of psychometrics, presented in terms of indices and statistics that may appear abstract or disconnected from everyday teaching (T. D. Brown et al., 2024; Estaji, 2024; Yongqi Gu & Lam, 2023). When data indicate that an item is too easy, for example, this is usually seen as a flaw in the test rather than as a clue about what students have mastered, perhaps due to effective instruction. Similarly, when distractors fail to attract students, this is interpreted merely as a weakness in item construction, rather than as evidence that students are using test-taking strategies—or that certain misconceptions are absent. In other words, test data can tell stories not only about the quality of the test itself, but also about the learning processes and instructional emphases that shape students' performance (Wallace & Ke, 2023; Yu & Xu, 2024).

By linking item analysis more directly to teaching and learning, educators can gain insights that go beyond statistics. Understanding why an item is too easy or why a distractor does not function can help teachers reflect on curriculum coverage, teaching methods, and student competencies (Wallace & Ke, 2023; Yu & Xu, 2024). This article takes the view that assessment and pedagogy are not separate silos but are deeply interconnected, and that test data should be treated as a valuable feedback loop for classroom practice.

Assessment in Educational Research

The relationship between assessment and instruction has long been emphasized in educational research. Hartati (2024) highlights how assessment can shape learning when feedback is meaningfully integrated into teaching; while most of this literature emphasizes classroom-based formative tasks, standardized or summative tests can also yield useful feedback if data are interpreted thoughtfully (T. D. Brown et al., 2024; Rejeki et al., 2023).

Item Analysis

Item analysis, as a technical procedure, has traditionally been situated within the field of psychometrics (T. D. Brown et al., 2024; Estaji, 2024; Yongqi Gu & Lam, 2023). Studies often focus on statistical properties such as item difficulty (p-value), discrimination index, and distractor efficiency. These measures are essential for ensuring the reliability and validity of tests. Yet, a growing body of scholarship suggests that item analysis can serve a pedagogical role. Lestari and Yusuf (2025) argue that language test performance reflects not only linguistic knowledge but also the instructional contexts that shape it. For instance, consistently high performance on certain item types may reflect that particular structures or skills have been emphasized in the curriculum. Conversely, persistent weaknesses may highlight gaps in instruction or misconceptions that require targeted remediation.

The Value of Distractor Analysis

Research has pointed to the interpretive value of distractor analysis. According to Semiun et al. (2022), non-functioning distractors are not only indicators of poor item quality, but may also suggest that students are not entertaining certain alternative conceptions, or that they have learned to quickly eliminate implausible options. This relates to the phenomenon of test-wiseness, where students apply strategies to answer correctly even without complete mastery of content (Rezkilaturahmi et al., 2025). Understanding how and why students choose particular distractors, then, opens a window into their reasoning and strategies, providing insights into both cognitive and instructional dimensions.

Pedagogical Implication on Test Data

More recent discussions on assessment literacy (Estaji, 2024; Lestari & Yusuf, 2025; Yu & Xu, 2024) emphasize the need for teachers to interpret test results not just as performance data but as pedagogical information. In this sense, item analysis can be seen as part of a broader movement to make

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

assessment more meaningful for teaching. Linking item statistics with curriculum content, teaching approaches, and learner profiles transforms test data into actionable knowledge.

In sum, while item analysis has traditionally been treated as a statistical exercise, there is increasing recognition that its real value lies in connecting numbers to narratives of learning. By situating item analysis within the context of teaching and curriculum, educators can better understand what test results reveal about students' competencies, instructional effectiveness, and areas for curriculum adjustment. This perspective not only bridges the gap between testing and teaching but also advances the broader goal of assessment-informed pedagogy in language education.

METHOD

Research Design

This study employed a quantitative descriptive design. Using a Classical Test Theory (CTT) framework, item analysis was conducted to examine the psychometric qualities of the Try-out test, including validity, item difficulty, discrimination indices, and distractor functioning. CTT was chosen because it provides well-established indices for systematically evaluating test items, thereby allowing researchers to identify problematic items, refine test instruments, and ensure fairness and reliability (Thompson & Ph, 2016). Item analysis und er CTT has been widely applied in educational measurement, as it highlights both the technical quality of items and the patterns of strengths and weaknesses that emerge from test-takers' responses (Bennett, 2017; Rohmatdi et al., 2024). Such analyses not only inform the revision of assessments but also provide valuable insights into teacher and learner competencies (Karim et al., 2021; Maharani & Putro, 2020; Sugianto, 2020). Therefore, item analysis was considered the most appropriate method for addressing the aims of this study.

Participants

This study involved 283 in-service English teachers of Indonesian junior, senior, and vocational high schools who were joining teacher professional development program (locally called as *PPG – Pendidikan Profesi Guru*) at Universitas Negeri Malang. This try-out aimed at preparing them before they joined a real UKPPPG test and predicting their better scores or results. All English teachers represented diverse backgrounds in terms of gender, region of origin, prior exposure to English learning, and varied years of experience (3–20 years). They also graduated from undergraduate degrees in English education or English literature. Participation was mandatory as part of the course assessment, and ethical clearance was obtained from the faculty research committee. Teachers' identities were anonymized during analysis.

Instrument

The analyzed test consisted of 35 PCK multiple-choice items, developed to measure English teachers' pedagogical, professional, social, and personal competences as presented the following blueprint.

Table 1. Blueprint of the PCK test

Type	Level			
	Easy	Medium	Difficult	Total
Simple MCQs	6 items	13 items	6 items	25 items
Complex MCQs	2 items	6 items	2 items	10 items
Total	8 items	19 items	8 items	35 items

Data Collection Procedure

The test was administered online via the university's learning management system (LMS). Teachers were given 90 minutes to complete the test. The LMS automatically recorded responses, scores, and completion times. Raw response data were exported into Excel, where each item was coded as 1 (correct) or 0 (incorrect).

Data Analysis

Item analysis was conducted using classical test theory (CTT) procedures (Ebel & Friesbie, 1991; Haladyna & Rodriguez, 2013; Hrich et al., 2024). Five indices were examined: item validity, discrimination, difficulty, distractor functioning, and item utilization. (1) *Item validity* was estimated through item–total correlations. Items with $r \geq 0.30$ were considered valid, while lower values indicated poor performance; (2) *Item discrimination* was computed using the 27% upper–lower group method. Items with a discrimination index of $D \geq 0.30$ were classified as acceptable, and those below were flagged for revision or removal; (3) *Item difficulty* was defined as the p-value, the proportion of correct responses. Items were categorized as *easy* (p ≥ 0.70), *moderate* (0.30 ≤ p ≤ 0.69), or *difficult* (p < 0.30); (4) *Distractor functioning* was evaluated based on the frequency of selection. Distractors chosen by fewer than 5% of test-takers were deemed non-functioning. Effective distractors were expected to attract low performers but not high performers; finally (5) *item utilization* decisions were made. Items meeting all criteria were *retained without revision*, those with minor flaws were *retained with revision*, and poorly performing items were *discarded*.

RESULTS AND DISCUSSION

Results

Descriptive Statistics of Test Performance

A total of 283 English teachers completed the 35-item English test. Table 2 summarizes the descriptive statistics.

Table 2. Descriptive Statistics of Total Scores (N = 283)

Statistics	Value
Mean	27.46
Standard Deviation	5.02
Minimum	6.00
Maximum	35.00
Skewness	−1.58
Kurtosis	3.44

As shown in Table 2, the Mean score is 27.46 and Standard Deviation is 5.02. This suggests that students averagely answered about 78% of the items correctly. The negatively skewed distribution indicates clustering of students at the higher end, with many scoring above 30. This pattern reflects the presence of several very easy items.

Item Validity

Item validity was measured using the item-total correlation. Out of 35 items, 16 items (46%) achieved acceptable validity (r ≥ .30) and 19 items (54%) fell below the acceptable threshold, rendering them invalid. This means that slightly fewer than half of the items were aligned with the overall construct of teacher competence being measured. Table 3 summarizes these findings.

Table 3. Item Validity Results

Category	Number of Items	Percentage
Valid	16	46%
Invalid	19	56%
Total	35	100%

The high proportion of invalid items suggests that many did not contribute meaningfully to the measurement of teacher competence, despite being developed according to the blueprint.

Reliability

The internal consistency of the test, as estimated by Cronbach's alpha, was α = .835, which is considered good and indicates that the items measure a coherent construct. This reliability supports the use of the test scores for both summative and diagnostic purposes.

Item Difficulty and Discrimination

The difficulty index (p) and discrimination indices were calculated for each item. Table 4 presents a summary of key results.

Table 4. Summary of Item Analysis (Selected Items)

Item	Difficulty (p)	Discrimination (r_it)	UL27 D	Interpretation
Q1	.99	.23	.04	Too easy, marginal discrimination
Q2	.87	.41	.34	Acceptable difficulty, good discrimination
Q12	.29	.25	.29	Difficult, marginal discrimination
Q20	.56	.34	.50	Moderate, good discrimination
Q23	.26	.19	.26	Too difficult, poor discrimination

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

Q35	.72	.33	.37	Moderate, good discrimination

Overall, 20 items fell within the acceptable difficulty range (.20–.90). However, 7 items were too easy ($p > .90$), including Q1, Q7, Q9, Q10, Q11, Q15, and Q16. These items contributed little to differentiation among students. Two items (Q12, Q23) were too difficult ($p < .30$). In terms of discrimination, about 40% of items showed good r_it (≥ .30), 30% were marginal (.20–.29), and the remaining 30% were poor (< .20). The Upper–Lower 27% index confirmed that items with higher r_it also separated high- and low-performing groups effectively.

Distractor Analysis

Distractor functioning revealed that several incorrect options attracted fewer than 5% of responses. For instance, in Q1, almost all students selected the correct answer, and distractors were virtually unused. This indicates a need for revision, as effective distractors should capture common misconceptions and provide diagnostic information about learning difficulties (Haladyna & Rodriguez, 2013). Across the test, about one-third of items had at least one non-functioning distractor.

To further interpret item performance, students were divided into three bands based on percentile cutoffs: the low band (≤ 26 points, ≤ 27th percentile), the middle band (27–31 points, 28th–72nd percentile), and the high band (≥ 32 points, ≥ 73rd percentile). This categorization allowed for a more nuanced analysis of how different proficiency groups approached the test items (Table 5).

Table 5. Proportion Correct by Performance Band (Selected Items)

Item	Low Band	Mid Band	High Band	UL27 D
Q12	.08	.32	.55	.47
Q20	.24	.61	.74	.50
Q23	.12	.29	.38	.26
Q35	.41	.68	.78	.37

As shown in Table 5, clear differences emerged across the three bands. Low-band students consistently struggled with higher-order items, particularly inference-based and grammar-in-context questions such as Q12 and Q23, where their proportion correct remained below .25. Middle-band students demonstrated partial competence, with proportions correct ranging between .29 and .68, but their performance across higher-order items lacked consistency. In contrast, high-band students performed strongly overall, achieving proportion correct scores above .55 on most items, although synthesis-oriented questions continued to present challenges. The discrimination indices (UL27 D) reinforced these findings. Items such as Q20 ($D = .50$) effectively differentiated between high- and low-performing students, indicating strong psychometric quality. Conversely, items such as Q23 ($D = .26$) showed weaker discrimination, suggesting the need for potential revision to better distinguish levels of student competence.

Discussion

This study analyzed 35 multiple-choice items from a try-out administered to in-service English teachers in Indonesia. By applying classical test theory, the analysis revealed significant weaknesses in item validity, discrimination, difficulty balance, and distractor functionality. These results provide insights into both the quality of the test and the competencies of the teachers. The discussion is structured around four key findings—validity, discrimination, difficulty, and distractors—followed by broader pedagogical and policy implications.

Validity and Teacher Competence Alignment

Validity is a cornerstone of test quality, ensuring that items measure intended constructs (Fulcher & Harding, 2021). With more than half of items found invalid, the test risked misrepresenting teachers' true competencies. Similar findings have been reported in Indonesian institutional assessments, where test items often lack construct validity due to superficial content alignment (Lestari & Yusuf, 2025). From a competency perspective, invalid items tended to test trivial or overly simplistic knowledge. Teachers' strong performance on such items may therefore reflect surface familiarity rather than mastery of the deeper competencies outlined in *Perdirjen 2626/B/HK/04.01/2023*. This echoes Yongqi Gu and Lam (2023) argument that validity lapses undermine the diagnostic value of assessments and limit their usefulness for professional growth.

Discrimination and Differentiation of Competencies

The ability of a test to discriminate between stronger and weaker candidates is critical in certification contexts (Ebel & Friesbie, 1991). Yet, 54% of items in this study had poor discrimination indices. This means that both high- and low-performing teachers were equally likely to answer correctly, reducing the test's effectiveness in evaluating readiness for certification. Research in similar Asian contexts reports the same issue: Vietnamese reading comprehension items showed low discrimination due to superficial distractors (Thu Thuy, 2023), while Malaysian teacher-made tests often failed to differentiate learners effectively (Sathasivam et al., 2023). In the Indonesian setting, poor discrimination suggests that items were insufficiently challenging or poorly designed to capture variation in teacher ability levels (Lestari & Yusuf, 2025). This aligns with Fulcher and Harding's (2021) observation that teacher assessments must not only certify but also differentiate competencies to target professional development.

Difficulty Distribution: The Skew Toward Easy Items

The overrepresentation of easy items (72%) contradicts the blueprint's mandate for a normal curve distribution. In the test design, easy items primarily measure lower-order skills such as recall and comprehension, while medium and difficult items target application, analysis, and evaluation (Bachman, 2022). The lack of medium and difficult items in the tryout limited its capacity to assess teachers' higher-order thinking skills (HOTS). This issue is consistent with prior research showing that Indonesian English assessments often privilege lower-order skills at the expense of analytical and evaluative competencies (Lestari & Yusuf, 2025). Similarly, Yongqi Gu and Lam (2023) found that Chinese national English exams overemphasized recall items, underrepresenting HOTS despite policy frameworks promoting critical thinking. By skewing toward easy items, the PPG tryout failed to provide an authentic measure of teachers' advanced professional competencies.

Distractor Functioning and Assessment Literacy

The widespread failure of distractors (74% non-functioning) is particularly alarming. Effective distractors are essential for multiple-choice validity, as they create plausible options that challenge weaker candidates without misleading stronger ones (Haladyna & Rodriguez, 2013). Non-functioning distractors lower both item difficulty and discrimination, leading to inflated scores and reduced reliability (Brown, 2019). This finding also points to a deeper issue: gaps in teachers' assessment literacy. As Yu and Xu (2024) argue, assessment literacy involves not only interpreting data but also designing items with effective distractors. Studies in Indonesia show that teachers often lack training in distractor construction, leading to tests with superficial options (Lestari & Yusuf, 2025). The current results suggest that both test developers and in-service teachers require targeted training in this area.

Competency Measured and Competency Missed

The four domains of teacher competence outlined in *Perdirjen 2626/B/HK/04.01/2023*—pedagogical, professional, social, and personal—provided a useful framework for interpreting the results of the item analysis. In terms of pedagogical competence, the test showed notable limitations. Poor item discrimination and the absence of more challenging items meant that teachers' ability to apply pedagogical principles in complex scenarios was not adequately captured. This finding aligns with Kusanagi (2022), who observed that Indonesian teacher assessments frequently fail to represent the nuanced realities of pedagogical practice. With regard to professional competence, teachers displayed strong performance on items assessing basic grammar and vocabulary, but higher-order competencies such as inference, synthesis, and academic English were underrepresented. Similar issues were highlighted by Sathasivam et al. (2023) in the Malaysian context, where assessments tended to emphasize surface-level knowledge rather than advanced language use.

For social and personal competencies, these domains were more explicitly targeted in the situational judgment test (SJT) section, which was not analyzed in the present study. Nevertheless, the weaknesses observed in pedagogical content knowledge (PCK) items suggest a broader issue of incomplete coverage across the four domains. Overall, while the test succeeded in measuring basic competencies, it fell short in capturing the higher-order professional and pedagogical skills that are increasingly recognized as essential for 21st-century teaching.

CONCLUSION

This study revealed significant weaknesses in the tryout test's psychometric quality. Less than half of the items were valid, over half failed to discriminate between high- and low-performing teachers, and the majority were too easy. Most distractors were non-functioning, and only a small fraction of items could be retained without revision. Together, these issues indicate that the test did not adequately capture the intended construct of teacher competence.

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

Despite these shortcomings, the results highlight important insights. Teachers demonstrated mastery of basic knowledge, but higher-order competencies—such as critical thinking, discourse analysis, and applied pedagogy—were under-assessed. This gap underscores the need to strengthen assessment literacy, both in test design and in classroom practice. Professional development should therefore focus on developing HOTS-oriented items, crafting effective distractors, and embedding item analysis into teachers' assessment routines.

Future research could extend the analysis to situational judgment items, apply Item Response Theory for more precise diagnostics, and link test data with classroom practice. Exploring AI-based tools for adaptive assessment also offers promising directions.

ACKNOWLEDGEMENT

This research was supported by a grant from the Teacher Professional Development Study Program, Sekolah Pascasarjana, Universitas Negeri Malang. The authors gratefully acknowledge this support, which made the completion of this study possible.

REFERENCES

Bachman, L. (2022). *Language Assessment for Classroom Teachers*. Oxford University Press.

Bennett, R. E. (2017). Advancing human assessment: The methodological, psychological and policy contributions of ETS. In *Methodology of educational measurement and assessment*. http://uml.idm.oclc.org/login?url=https://search.proquest.com/docview/2008438806?accountid=14569%0Ahttp://primo-pmtna01.hosted.exlibrisgroup.com/openurl/UMB/01UMB_INST/umb_services_page??url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:book&genre=book

Brown, H. D. (2019). *Language Assessment: Principles and Classroom Practices*. Pearson Education.

Brown, T. D., Barnes, M., & Finefter-Rosenbluh, I. (2024). Teacher perspectives and experiences of assessment literacy in Victorian junior secondary schools. *Australian Journal of Education*, *68*(1), 5–22. https://doi.org/10.1177/00049441231214022

Ebel, R. L., & Friesbie, D. A. (1991). Essentials of Educational Measurem. *Journal of Educational Measurement*, 1–370. https://ebookppsunp.files.wordpress.com/2016/06/robert_l-ebel_david_a-_frisbie_essentials_of_edbookfi-org.pdf

Estaji, M. (2024). Perceived need for a teacher education course on assessment literacy development: insights from EAP instructors. *Asian-Pacific Journal of Second and Foreign Language Education*, *9*(1). https://doi.org/10.1186/s40862-024-00272-2

Fulcher, G., & Harding, L. (2021). The Routledge Handbook of Language Testing. In *The Routledge Handbook of Language Testing*. https://doi.org/10.4324/9781003220756

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. Routledge.

Hartati, N. (2024). Item analysis of pre-service EFL teachers' formative test in teaching practicum for program evaluation. *LingTera*, *11*(2), 145–158. https://doi.org/10.21831/lt.v11i2.67170

Hrich, N., Azekri, M., & Khaldi, M. (2024). Artificial Intelligence Item Analysis Tool for Educational Assessment: Case of Large-Scale Competitive Exams. *International Journal of Information and Education Technology*, *14*(6), 822–827. https://doi.org/10.18178/ijiet.2024.14.6.2107

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

Kusanagi, K. N. (2022). Importing and Exporting Lesson Study. In *Education in the Asia-Pacific Region* (Vol. 69). https://doi.org/10.1007/978-981-19-5928-8_3

Lestari, S., & Yusuf, F. N. (2025). Aligning Assessment Practices with Learning Objectives: A Case of EFL Classes in Indonesia. *IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)*, *10*(1), 145. https://doi.org/10.21093/ijeltal.v10i1.1973

Maharani, A. V., & Putro, N. H. P. S. (2020). Item Analysis of English Final Semester Test. *Indonesian Journal of EFL and Linguistics*, *5*(2), 491. https://doi.org/10.21462/ijefl.v5i2.302

Rejeki, S., Bayu, A., Sari, P., Iswahyuni, D., Yogyanti, D. W., Sutanto, S., & Anggia, H. (2023). Discrimination index, difficulty index, and distractor efficiency in MCQs English for academic purposes midterm test. *Journal of English Language and Pedagogy Journal of English Language* *and Pedagogy Journal of English Language and Pedagogy*, *6*(1), 1–11.

Rezkilaturahmi, Istiqlal, M., Pancoro Setyo Putro, N. H., Istiyono, E., & Widihastuti. (2025). Evaluating English Language Test Items Developed by Teachers: An Item Response Theory Approach. *Voices of English Language Education Society*, *9*(1), 218–230. https://doi.org/10.29408/veles.v9i1.27644

Rohmatdi, A., Martono, M., Zafrullah, Z., & Safitri, R. (2024). Optimizing Question Quality in Junior High School Exams: Classical Test Theory Evaluation with ITEMAN 4.3. *AL-ISHLAH: Jurnal Pendidikan*, *16*(4), 4363–4375. https://doi.org/10.35445/alishlah.v16i4.5671

Sathasivam, R. V., Samuel, M., Nor, N. M., Tee, M. Y., & Leong, K. E. (2023). Assessment for learning: Espoused and enacted practices of Malaysian teachers. *Pertanika Journal of Social Sciences and Humanities*, *27*(T2), 47–62.

Semiun, T. T., Wisrance, M. W., & Napitupulu, M. H. (2022). English Summative Test: The Quality of Its Items. *English Education:Journal of English Teaching and Research*, *7*(2), 119–127. https://doi.org/10.29407/jetar.v7i2.18347

Sugianto, A. (2020). Item analysis of English summative test: EFL teacher-made test. *Indonesian EFL Research and Practices*, *1*(1), 35–54. https://journal.iaima.ac.id/i-efl/article/view/4

Thompson, N. A., & Ph, D. (2016). *Thompson_2016_Classical_Test_Theory_with_CITAS*.

Thu Thuy, N. T. (2023). Vietnamese High School Students' Perceptions And Practices Of Metacognitive Reading Strategies: A Case Study. *International Journal of Scientific and Research Publications*, *13*(6), 92–98. https://doi.org/10.29322/ijsrp.13.06.2023.p13814

Varier, D., Powell, M. G., Dodman, S., Ives, S. T., DeMulder, E., & View, J. L. (2024). Use and Usefulness of Assessments to Inform Instruction: Developing a K-12 Classroom Teacher Assessment Practice Measure. *Educational Assessment*, *29*(1), 36–57. https://doi.org/10.1080/10627197.2024.2316907

Wallace, M. P., & Ke, H. (2023). Examining the Content Alignment Between Language Curriculum and a Language Test in China. *Teflin Journal*, *34*(1), 116–135. https://doi.org/10.15639/teflinjournal.v34i1/116-135

Yongqi Gu, P., & Lam, R. (2023). Developing Assessment Literacy for Classroom-Based Formative Assessment. *Chinese Journal of Applied Linguistics*, *46*(2), 155–161. https://doi.org/10.1515/CJAL-2023-0201

Yu, C., & Xu, W. (2024). Writing assessment literacy and its impact on the learning of writing: A netnography focusing on Duolingo English Test examinees. *Language Testing in Asia*, *14*(1). https://doi.org/10.1186/s40468-024-00297-x

ISSN 2963-7856 (print), ISSN 2961-9890 (online)