BEYOND COMPARISON: HIGHLIGHTING THE GAP SIMILARITIES IN DUOLINGO AND ENGLISHSCORE TESTS

Samsul Khabib[1], Rarasaning Satianingsih[2], Nur Rohmah[3]
[1]English Language Education Department – Universitas PGRI Adi Buana Surabaya, Indonesia
[2]Primary Teacher Education Department – Universitas PGRI Adi Buana Surabaya, Indonesia
[3]Indonesian Language Education Department – Universitas PGRI Adi Buana Surabaya, Indonesia
[1]Doctoral Student, English Language Education Department – Universitas Katolik Widya Mandala Surabaya, Indonesia

Abstract. *The Duolingo English Test (DET) and the British Council's EnglishScore are gaining popularity as alternative digital English tests. Previous research has tended to compare the strengths of both platforms, but has rarely identified patterns of weaknesses that users may share, albeit with varying degrees of severity. This study aims to explore and thematically compare the weaknesses reported by DET and EnglishScore users in the Indonesian EFL context. The study used a qualitative approach with thematic analysis (Braun & Clarke, 2006) of open-ended comments from two independent surveys. A total of 129 DET users and 133 EnglishScore users (total N=262) – Teacher Professional Education program students at a private university in Surabaya – completed a user experience questionnaire, including open-ended questions about the most difficult parts to understand. The analysis yielded six themes of weaknesses that emerged across both platforms: (1) unclear or confusing instructions; (2) technical stability issues (crashes/self-exiting); (3) audio/speaking issues (voice detection and speed); (4) too short a turnaround time; (5) difficulty for English beginners; and (6) lack of specific feedback. Although the frequency of complaints was significantly higher on DET, the pattern of weaknesses experienced was qualitatively the same across both platforms. Digital English language testing platforms, regardless of their performance level, face similar functional challenges. Developers need to prioritize improvements in instruction clarity, technical stability, and accessibility for users with low English proficiency. This research contributes to the literature on user experience evaluation of digital assessment platforms in developing countries.*

Keywords: *digital English test; Duolingo English Test; EnglishScore; platform weaknesses; thematic analysis; user experience; Indonesian EFL.*

Introduction

The digital revolution has fundamentally transformed the landscape of English language assessment. The COVID-19 pandemic catalyzed this transformation, as the closure of test centers worldwide forced educational institutions to seek viable remote testing alternatives (Isbell & Kremmel, 2020 in Isaacs et al., 2023). In this context, digital English testing platforms such as the Duolingo English Test (DET) and the British Council's EnglishScore have emerged as promising solutions, offering flexibility, accessibility, and lower costs than traditional tests like the IELTS or TOEFL iBT (Isaacs et al., 2023).

The DET, a fully automated, computer-based adaptive test, has been adopted by thousands of institutions worldwide (Isaacs et al., 2023). A large-scale study by Burstein

DOI: 10.5281/zenodo.19781170

et al. (2025) of 25,969 DET participants showed that access to AI-based practice tests can improve confidence and performance. On the other hand, EnglishScore, developed by the British Council, offers a fast and accurate mobile-based assessment for measuring grammar, vocabulary, reading, and listening (Arisandi et al., 2025).

However, behind this enormous potential, significant challenges emerge that are often overlooked in academic discussions that tend to focus on comparative excellence. Arisandi et al. (2025) revealed that test takers' initial experiences with EnglishScore were characterized by multidimensional difficulties: time pressure, limited vocabulary, text complexity, and unfamiliarity with digital formats and test strategies. In the Indonesian context, access to standardized tests is limited and English proficiency is generally low. Maknun et al. (2024)reported that Indonesian EFL teachers face obstacles in technology-based assessment, including lack of time, large student numbers, and inadequate knowledge and training.

This research starts from the premise that focusing solely on comparing the strengths of digital platforms ignores an important reality: both platforms may face the same types of weaknesses, albeit with varying degrees of severity. By understanding the patterns of weaknesses that emerge across both platforms, this research contributes to the development of more inclusive and accessible digital testing platforms, particularly in the Indonesian EFL context.

Based on the above background, this study was proposed to answer three research questions, namely: what are the themes of weaknesses that emerge in the user experience of DET and EnglishScore, what are the patterns of similarities in weaknesses between the two platforms, and what are the implications of these weakness patterns for the development of better digital test platforms, especially in the Indonesian EFL context. In line with these formulations, this study aims to identify the themes of weaknesses reported by DET and EnglishScore users, compare and analyze the patterns of similarities in weaknesses between the two platforms, and formulate recommendations for platform developers and other stakeholders.

More broadly, this research is expected to provide theoretical and practical benefits. Theoretically, this research contributes to the literature on technology-based language assessment by shifting the focus from comparing advantages to identifying patterns of shared weaknesses, an approach that aligns with the calls of Chuang & Yan, (2025) and Aryadoust et al., (2024) regarding the dangers of adopting technology without understanding its challenges (in Shvetsova et al., 2025). Furthermore, this research also enriches the discussion on factors that influence user experience in digital assessment (Chikezie et al., 2025). Practically, the findings of this study can provide input for platform developers (Duolingo and the British Council) to improve features and interfaces; for educational institutions to consider choosing a platform that suits their needs and is aware of their respective weaknesses; for test takers to provide a realistic picture of the challenges they may face; and for future researchers to provide a foundation for further study on platform effectiveness and the development of more inclusive assessments.

LITERATURE REVIEW

Theoretical basis

This research is based on three theoretical pillars: technology-based language assessment, user experience, and functional weaknesses of digital platforms in the EFL context.

Automated assessments have developed rapidly, offering ease of access, flexibility, and lower costs than conventional tests (Isbell & Kremmel, (2020) in Isaacs et al. (2023). However, challenges regarding validity, reliability, and fairness remain. Automated assessments of productive skills (writing, speaking) face reliability issues due to the complexity of assessing subjective language expression (review of Automated Language Assessment, 2025).

This is crucial because the success of a platform is determined not only by technical accuracy but also by user comfort and confidence. Research at BRAC University (2025) identified three main barriers: psychological, cognitive, and technological (Gkintoni et al., 2025). Sun et al. (2025) added that emotional factors (anxiety, distrust, stress) are the most significant barriers to the adoption of AI language assessment.

Digital platforms are cross-platform. Arisandi et al. (2025)found five categories of difficulty with EnglishScore: time pressure, limited vocabulary, text complexity, translation challenges, and unfamiliarity with the test. Research on digital assessment development (2025) highlights that digital technology tends to ignore subjectivity and individual differences, and poses ethical risks.

Duolingo English Test (DET) and British Council EnglishScore

DET is a high-stakes, computer-based adaptive test that measures reading, writing, listening, and speaking in approximately 45 minutes at a low cost (Pearson, 2023). The DET has been adopted by thousands of institutions, but faces concerns regarding the fairness, security, and authenticity of the test items (Kang et al., 2024). Isaacs et al. (2023) found that students who received the DET had lower academic success than those who received the IELTS/TOEFL iBT. In Indonesia, Wali et al. (2025) reported that 96.5% of students had never used the DET for official testing, with scores still below the requirements of international mobility programs.

EnglishScore is a mobile-based test from the British Council that measures grammar, vocabulary, reading, and listening (Kurniati et al., 2025). Results can be obtained in minutes. Ulinuha et al. (2025) reported significant improvements in scores after a brief intervention, but the effects were modest. Arisandi et al. (2025) revealed that test takers experienced multidimensional difficulties, including time pressure and cognitive overload.

Challenges of Digital Assessment in the Indonesian EFL Context

Indonesia faces unique challenges: generally low English proficiency, uneven access to digital infrastructure, and limited digital literacy. Maryansyah & Danim, (2024) reported that Indonesian EFL students experienced technical difficulties during online exams, were dissatisfied with the lack of feedback, and were concerned about academic honesty. Research on technology integration in EFL assessments also highlighted limited internet access and unstable networks.

The specific challenges that emerged with DET and EnglishScore showed similar patterns: (1) clarity of instructions – unfamiliarity with digital formats led to strategic disorientation Arisandi et al. (2025); Kang et al. (2024); (2) time pressure – tight constraints were a major difficulty (Arisandi et al., 2025); (3) technical stability – connection issues and system failures were common (Maryansyah & Danim, 2024); (4) anxiety and self-confidence – emotional factors hindered AI adoption (Sapru, 2026).

State of the Art and Research Originality

Previous research has tended to focus on the validity, reliability, and predictive power of the DET compared to the IELTS/TOEFL, or on the specific difficulties of the EnglishScore. The research gaps identified are: (1) no research has explicitly compared functional weakness patterns between the DET and the EnglishScore; (2) most research focuses on user strengths rather than weaknesses; (3) qualitative thematic approaches to identifying shared weakness patterns are still rare; (4) the Indonesian context has not been widely explored in comparative studies of digital platforms.

This research fills this gap by shifting from a comparative paradigm of strengths to identifying shared patterns of weaknesses across two different platforms in the Indonesian EFL context. This approach provides a new perspective on digital assessment evaluation and provides practical contributions for platform developers.

Research Thinking Framework

This research starts from the premise that evaluating digital assessment platforms is not sufficient by simply measuring technical excellence (validity, reliability, predictiveness) or overall satisfaction. User experience—particularly in the Indonesian EFL context—presents a crucial, often overlooked dimension: functional weaknesses that are shared, albeit with varying degrees of severity. Using thematic analysis of open-ended comments from DET and EnglishScore users, this research will identify emerging themes of weaknesses across both platforms, analyze their patterns of similarity, and formulate recommendations for platform developers and other stakeholders. Focusing on shared weaknesses—rather than simply differences in strengths—is expected to contribute meaningfully to the development of more inclusive and user-friendly digital assessments in developing countries like Indonesia.

RESEARCH METHODOLOGY

Research Design

This research employed a qualitative approach with a comparative case study design. This design was chosen because the primary objective of the study was to understand and compare in-depth patterns of impairment experienced by users of two digital English language testing platforms (DET and EnglishScore) in their natural contexts. The qualitative approach allows for a holistic exploration of participants' meanings and perspectives (Creswell, 2013), while the comparative case study design facilitates the identification of similarities and differences across cases (Yin, 2018).

Research Participants

Participants were students of the Teacher Professional Education program (Pendidikan Profesi Guru/ PPG) at a private university in Surabaya who participated in a survey of DET and EnglishScore user experiences. 129 respondents completed the DET and 133 for EnglishScore. The characteristics of both groups were relatively balanced: gender (48.92% male, 51.08% female), average age 22–28 years, the majority used Android devices (70–74%), and the majority had never used the tested platforms (87–91%). All participants signed a voluntary informed consent.

Data Sources and Instruments

The study used secondary data in the form of open-ended comments from an online questionnaire. The question analyzed was: “Give 1-2 sentences of advice or your impressions regarding the most difficult part of the platform you tried earlier.” Comment data is found in the DET survey files (pages 70–72) and EnglishScore (pages 70–72). Total relevant comments: 112 from DET (86.8%) and 118 from EnglishScore (88.7%).

The main instrument was the researcher herself (human as instrument) assisted by a thematic analysis guide, coding sheet, reflection journal, and audit trail.

Data Analysis Techniques

Data analysis used Braun & Clarke, (2006) six-step thematic analysis: (1) familiarization by repeatedly reading all comments; (2) inductively creating initial codes (e.g., "time too short," "microphone indicator missing"); (3) grouping codes into potential themes; (4) reviewing themes to ensure internal coherence and external distinction; (5) defining and naming themes accurately; (6) writing a report supported by direct quotations.

Data Validity

The research applied Lincoln & Guba, (1985) trustworthiness criteria, adapted by (Creswell, 2013). Credibility was achieved through sustained engagement with the data, source triangulation (comparing DET and EnglishScore comments), peer debriefing, and negative case analysis. Transferability was achieved through rich and detailed descriptions of the context and participants. Dependability was achieved through documented audit trails throughout the analysis process. Confirmability was achieved through reflecting on personal biases and transparently presenting original quotations.

Research Ethics

The research has met ethical principles: informed consent from all participants, confidentiality of identity (names are not published, data is stored on a password-protected computer), the right to withdraw without penalty, and permission to use secondary data from the original researchers.

Methodological Limitations

The researchers acknowledged several limitations. First, secondary data prevented them from asking clarifying questions to participants. Second, in the DET group, there was no explicit question confirming that participants had actually completed the DET (unlike the EnglishScore), thus risking misclassification. Third, the homogeneous sample (PPG students from one university) limits generalizability to the broader Indonesian EFL population. Fourth, the short comments (1-2 sentences) limited the depth of information compared to in-depth interviews. These limitations were transparently acknowledged and serve as reflections for future research.

RESULTS AND DISCUSSION

Comment Data Overview

Of the 129 DET respondents, 112 (86.8%) provided relevant comments; of the 133 EnglishScore respondents, 118 (88.7%) provided relevant comments. Negative comments (complaints/identification of difficulties) were significantly more prevalent on DET, while EnglishScore was mostly positive or neutral. However, for the thematic analysis of weaknesses, focus was placed on comments identifying difficulties, confusion, or technical challenges on both platforms.

Themx Weaknesses on Both Platforms

Thematic analysis yielded six themes of weakness that emerged in both the DET and EnglishScore, albeit with varying frequencies. The following table summarizes the comparison of representative intensities:

Weakness Theme	DET Frequency	EnglishScore Frequency
Unclear/confusing instructions	High (≥40%)	Low (~10%)

Technical stability issues (crash/self-exit)	Medium (~25%)	Low (~8%)
Audio/speaking issues	High (~35%)	Low (~5%)
The processing time is too short	Medium (~20%)	Very low (~3%)
Difficulties for English beginners	Medium (~15%)	Low (~7%)
Lack of specific feedback	Low (~10%)	Very low (~2%)

The following is an explanation of each theme accompanied by representative quotes.

Theme 1: Instructions are unclear or confusing

DET: “The instructions on what to do when there is a question that must be clarified are somewhat unclear.” (p. 70)

EnglishScore: "The reading instructions were unclear. When I tried to repeat them, I ended up starting from the beginning." (p. 70)

Both platforms struggled to deliver intuitive instructions. The DET had more systemic problems, perhaps due to the adaptive format; EnglishScore was limited to specific features (drag-and-drop, room scanning). This finding is consistent with Arisandi et al.'s (2025) study of strategic disorientation and Kang et al. (2024) study of item authenticity.

Theme 2: Technical stability constraints (crash/self-exit)

DET: “Suddenly, he came out on his own and started the test again from the beginning, even though he had previously reached stage 6 or 7.” (p. 71)

EnglishScore: “I often have signal problems and suddenly exit the platform by myself.” (p. 71)

Both platforms are susceptible to fluctuating internet connection quality, but EnglishScore has a better recovery mechanism, as restart complaints are more prevalent in DET. This is in line with Maryansyah & Danim (2024) regarding the technical difficulties of online assessments in Indonesia.

Theme 3: Audio/speaking issues (voice detection & speed)

DET: “There is no indication whether the sound has been recorded or not.” “The sound is too fast.” (pp. 71-72)

EnglishScore: “A slight cough has been detected.” “For the listening section, the clarity of the voice has been improved.” (p. 70)

DET was very weak in visual microphone indication, causing anxiety; EnglishScore was better, but there were still complaints about speed and sensitivity. This reinforces Sapru (2026) findings on emotional factors in AI adoption.

Theme 4: The processing time is too short

DET: “The time given for each question is not long enough.” “Please give me a little more time.” (p. 70)

EnglishScore: “Give more time.” “Time duration speed.” (p. 71)

Time pressure was a major difficulty identified by Arisandi et al. (2025). The DET is more intense due to its adaptive format and longer test duration (45-60 minutes vs. 30-40 minutes for EnglishScore).

Theme 5: Difficulties for English beginners

DET: "All commands are in English, so beginners may have a bit of difficulty understanding them if they are not translated." (p. 70)

EnglishScore: "For those who can speak English it's easy, for those who can't it's difficult." "Guide to learning Indonesian." (pp. 70, 72)

Monolingual English instruction creates a double burden for beginners. Platforms need to consider local language guidance for technical instructions without compromising the validity of the test content.

Theme 6: Lack of specific feedback

DET: "At the end it is not explained in detail which parts need to be fixed." (p. 71)

EnglishScore: "Each question is given knowledge of right and wrong and its explanation. To make it easy to learn." (p. 70)

Both platforms only provide a final score without any subskill analysis. Yet, good feedback is key to the effectiveness of practice tests (Burstein et al., 2025).

Comparison of Intensity and Interpretation

Quantitatively, EnglishScore outperformed DET in every aspect of weaknesses (the frequency of complaints was significantly lower). However, qualitatively, the pattern of weaknesses experienced was the same on both platforms: confusing instructions, crashes, audio issues, time pressure, beginner accessibility, and lack of feedback.

The similarity of patterns is due to the fundamental similarities of both platforms: digital format (dependence on connection, camera, microphone), AI-based automated assessment (less transparent), global target audience (monolingual English instruction), and time pressure for efficiency.

The difference in intensity (DET is more severe) is explained by: (1) DET does not have a visual microphone indicator, (2) the recovery mechanism from a bad connection break (restart from scratch), (3) the adaptive format is more complex, (4) the test duration is longer.

Implications for Platform Development

Developers need to prioritize fixing common weaknesses: (1) real-time visual indicator for microphone; (2) automatic progress saving when connection drops; (3) audio speed (slow) option; (4) local language guides for technical instructions; (5) per-subskill feedback; (6) practice mode without time pressure.

Summary of Findings

This chapter presents six themes of weaknesses that emerged across both platforms. The key finding: While EnglishScore is quantitatively superior, the patterns of weaknesses experienced by users are essentially the same. This indicates that challenges in digital language assessment are fundamental and cross-platform, not solely due to the quality of technical implementation.

CONCLUSION AND SUGGESTIONS

Conclusion

This study identifies similar patterns of weaknesses in two digital English language testing platforms, the Duolingo English Test (DET) and the British Council's EnglishScore, based on user experiences in the Indonesian EFL context. Through a thematic analysis of open-ended comments from 129 DET users and 133 EnglishScore users, six key conclusions can be drawn.

First, six themes of weaknesses were found that emerged across both platforms: (1) unclear or confusing instructions, (2) technical stability issues (crashes/self-exiting), (3) audio/speaking issues (voice and speed detection), (4) too short a timeframe, (5) difficulties for English beginners, and (6) lack of specific feedback. These six themes indicate that functional challenges in digital language assessment are cross-platform.

Second, qualitatively, the patterns of difficulties experienced by DET and EnglishScore users were similar, although with varying frequency and severity. EnglishScore consistently showed lower levels of complaint intensity across all themes, but the types of difficulties reported (confusion with instructions, concerns about voice detection, frustration with app exits) were essentially similar.

Third, the most dominant weaknesses in the DET were confusing instructions and audio/speaking issues (particularly the lack of a voice recording indicator), while in the EnglishScore, weaknesses were sporadic and related to specific features (drag-and-drop, room scanning). These differences in intensity were due to interface design, error recovery mechanisms, format complexity, and test duration.

Fourth, these findings confirm that focusing on comparing platform strengths is insufficient. Users face fundamentally similar barriers regardless of platform. Improvement efforts should be directed at shared, systemic weaknesses.

Fifth, in the context of Indonesian EFL, the lack of accessibility for beginners and the reliance on a stable internet connection are particularly relevant challenges. Platforms need to consider the real-world conditions of users in developing countries.

Sixth, this research successfully answered the problem formulation, provided theoretical contributions (paradigm shift from comparing advantages to identifying common weaknesses), and practical (recommendations for improvements for developers).

Suggestion

For platform developers (Duolingo and the British Council): (1) add a real-time visual indicator for the microphone; (2) implement an automatic progress saving mechanism when the connection is lost; (3) provide an audio speed (slow) option for listening; (4) provide local language guidance for technical instructions; (5) provide specific per-subskill feedback; (6) provide a practice mode without time pressure.

For educational institutions: Select a platform according to infrastructure readiness, provide a brief orientation about the test flow and technical tips (microphone, headset, stable connection), and provide a technical helpdesk during implementation.

For test takers: Do technical preparation (storage space, battery, connection, headset), use practice mode to understand the question format, and manage your time wisely.

For further researchers: conduct in-depth interviews or FGDs, direct observation (screen recording), larger-scale quantitative studies, comparisons with other platforms (TOEFL iBT Home Edition, IELTS Indicator), and intervention experiments to test the effectiveness of the proposed solutions.

Research Limitations

The researchers acknowledged several limitations: (1) secondary data, so clarification or probing could not be performed; (2) in the DET group there was no explicit confirmation that participants actually completed the DET; (3) a homogeneous sample (PPG students from one university) limited generalizability; (4) brief comments

(1-2 sentences) limited the depth of information; (5) did not measure the impact of weaknesses on final scores or admission decisions.

Closing

This research shows that despite differences in technical performance and satisfaction levels, DET and EnglishScore face similar weaknesses. Educational technology evaluations should not simply ask "which platform is better?" but also "what do they both find difficult for users?" By understanding shared weaknesses, developers, institutions, and researchers can collaborate to create more inclusive, accessible, and user-friendly digital language assessments.

References:

1.Arisandi, V., Yulianti, H. T., Prihamdani, D., Juanda, J., & Mulya, P. W. (2025). Test-taking difficulties in Englishscore reading assessment: Insights from learner experiences. *TELL-US JOURNAL*, *11*(3), 900–917. https://doi.org/10.22202/tus.2025.v11i3.10027

2. Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence*, *6*, 100204. https://doi.org/10.1016/j.caeai.2024.100204

3. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

4. Burstein, J., Cardwell, R., Chuang, P.-L., Michalowski, A., & Nydick, S. (2025). *Exploring AI-Enabled Test Practice, Affect, and Test Outcomes in Language Assessment* (arXiv:2508.17108). arXiv. https://doi.org/10.48550/arXiv.2508.17108

5. Chikezie, C., Chanpaisaeng, P., Agarwal, P., Afroz, S., Madhwani, B., Choudhuri, R., Anderson, A., Velhal, P., Morreale, P., Bogart, C., Sarma, A., & Burnett, M. (2025). Measuring SES-related traits relating to technology usage: Two validated surveys. *Empirical Software Engineering*, *30*(6), 159. https://doi.org/10.1007/s10664-025-10683-5

6. Chuang, P.-L., & Yan, X. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. *System*, *134*, 103846. https://doi.org/10.1016/j.system.2025.103846

7. Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (third edition). SAGE.

8. Gkintoni, E., Antonopoulou, H., Sortwell, A., & Halkiopoulos, C. (2025). Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in Redefining Learning Efficacy. *Brain Sciences*, *15*(2), 203. https://doi.org/10.3390/brainsci15020203

9.Isaacs, T., Hu, R., Trenkic, D., & Varga, J. (2023). Examining the predictive validity of the Duolingo English Test: Evidence from a major UK university. *Language Testing*, *40*(3), 748–770. https://doi.org/10.1177/02655322231158550

10. Isbell, D. R., & Kremmel, B. (2020). Test Review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *37*(4), 600–619. https://doi.org/10.1177/0265532220943483

11. Kang, O., Yan, X., Kostromitina, M., Thomson, R., & Isaacs, T. (2024). Fairness of using different English accents: The effect of shared L1s in listening tasks of the Duolingo English test. *Language Testing*, *41*(2), 263–289. https://doi.org/10.1177/02655322231179134

12. Kurniati, D., Riyono, A., Solaeman, A., R. Millan, A., & Noordin, N. (2025). Exploring English proficiency levels of Indonesian university students through EnglishScore-based assessment. *Indonesian EFL Journal*, *11*(3), 539–552. https://doi.org/10.25134/9kjdz259

13. Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry* (Nachdr.). Sage.

14. Maknun, L., Zamzani, Z., & Jamilah, J. (2024). Unveiling Indonesian EFL Teacher's Perceptions and Challenges of Technology-based Assessment as and for Learning. *International Journal of Language Testing*, *14*(1). https://doi.org/10.22034/ijlt.2023.400628.1260

15. Maryansyah, Y., & Danim, S. (2024). Experiences, Perceptions, and Challenges of Indonesian EFL University Students with Online Assessment in the Digital Age. In M. Kristiawan, N. D. Lestari, D. Samitra, Z. F. Rozi, M. N. Naser, R. M. Valianti, M. Muthmainnah, B. Badeni, F. A. Yanti, D. Apryani, O. L. Agusta, J. Siska, E. Viona, E. Purwandari, & R. D. Riastuti (Eds.), *Online Conference of Education Research International (OCERI 2023)* (Vol. 775, pp. 719–732). Atlantis Press SARL. https://doi.org/10.2991/978-2-38476-108-1_71

16. Pearson, W. S. (2023). Test review: High-stakes English language proficiency tests—Enquiry, resit, and retake policies. *Language Testing*, *40*(4), 1022–1035. https://doi.org/10.1177/02655322231186706

17. Sapru, A. (2026). Psychological resistance to AI: How regulatory focus fuels AI anxiety and negative attitudes toward AI. *Technology in Society*, *86*, 103251. https://doi.org/10.1016/j.techsoc.2026.103251

18. Shvetsova, I., Lytvynska, S., Davydova, T., Romanchuk, S., & Rusavska, O. (2025). Intelligent Language Systems and Technical Linguistic Solutions in the Digital Environment. *International Journal on Culture, History, and Religion*, *7*(SI1). https://doi.org/10.63931/ijchr.v7iSI1.195

19. Sun, X., Dou, W., & Yang, Y. (2025). The socio-emotional dangers of using Artificial Intelligence (AI) technologies in second language (L2) education: Unveiling Chinese EFL teachers' perceptions and experiences. *Acta Psychologica*, *261*, 105956. https://doi.org/10.1016/j.actpsy.2025.105956

20. Ulinuha, A., Parnawati, T. A., & Chotimah, C. (2025). From perception to performance: Investigating students' learning outcomes in mobile-based English proficiency testing. *Inteligensi : Jurnal Ilmu Pendidikan*, *8*(1), 44–53. https://doi.org/10.33366/ilg.v8i1.7003

21. Wali, A. I., Syarif, A. R., & Syahriani, R. (2025). *Proficiency level score of Duolingo English practice test*. *11*(2).

22. Yin, R. K. (2018). *Case study research and applications: Design and methods* (Sixth edition). SAGE.