

User Experience in Digital English Assessment: An Exploratory Comparison
of Duolingo English Test and EnglishScore
Samsul Khabib¹, Las Johansen B. Caluza²

¹Universitas PGRI Adi Buana Surabaya, Jl. Dukuh Menanggal XII Surabaya, East Java 60234, Indonesia

² Leyte Normal University (LNU), P. Paterno St., Brgy. 50, Tacloban City, Leyte 6500, Philippines

Corresponding email: samsulkhabib@unipasby.ac.id

ARTICLE INFO	ABSTRACT
Article history: Received: February 10th, 2026 Revised: February 27th, 2026 Accepted: April 25th, 2026 Published: April 30th, 2026 Keywords: Digital language assessment, User experience, Duolingo English Test, EnglishScore, Teacher Educational, Education,	Digital language assessments expand access and speed up the reporting of results, but user experience can introduce variance that is unrelated to language constructs. This study compared user experiences of the Duolingo English Test (DET) and EnglishScore among Indonesian Professional Teacher Education (PPG) participants. A convergent mixed-methods design was used, combining a five-point Likert-scale survey and open-ended comments. The analysis included 264 completed responses, with 130 responses for the DET and 134 for EnglishScore. Twenty-five survey items were organized into five dimensions: access and ease of use, clarity of instructions, confidence in the score, usefulness of feedback, and technical readiness. Quantitative data were analyzed descriptively, and open-ended comments were examined thematically. The overall mean for EnglishScore ($M = 4.34$) was higher than for the DET ($M = 3.99$) across all dimensions. The largest gap was found in clarity of instructions (4.31 versus 3.89), especially in understanding the test sections. Participant comments highlighted four main sources of difficulty: instructions and navigation, time pressure, audio or recording issues, and connection or device glitches. Bringing together the findings suggests that technical accessibility alone is not enough; procedural clarity and the usability of score reports also shape the assessment experience. Since the two platforms were assessed by different groups and individual data are not available, these results are not meant to provide evidence of psychometric superiority for either test. The findings support the use of pre-test orientations, device checks, interactive sample tasks, and more diagnostic result reports when implementing digital assessments in EFL contexts. <i>This is an open access article under the CC BY-SA license.</i> 
How to cite: Khabib, S., & B. Caluza, L. J. (2026). User Experience in Digital English Assessment: An Exploratory Comparison of Duolingo English Test and EnglishScore. English Language Teaching Methodology, 6(1), 52–64. https://doi.org/10.56983/eltm.v6i1.2046	

Corresponding Author:

Samsul Khabib

English Language Education Department,

Universitas PGRI Adi Buana Surabaya

Jl. Dukuh Menanggal XII No.12 Surabaya, Eats Java, Indonesia

Email: samsulhabib@unipasby.ac.id

INTRODUCTION

Digitization has changed language assessment from simply moving questions onto a screen into an ecosystem that brings together task presentation, adaptive measurement, response recording, remote proctoring, automated scoring, and results reporting. Technology can widen access, speed up administration, and allow for new task formats. Still, technology can also bring challenges beyond language proficiency, such as unclear interfaces, device limitations, unstable connections, or unfamiliarity with test procedures (Chapelle & Douglas, 2006; Douglas, 2007). Within a validity framework, these conditions matter because they can introduce variance unrelated to the construct being measured. The DET and EnglishScore represent two different digital assessment models. The DET is an adaptive, computer-based English language proficiency test given online with remote proctoring; its technical documentation describes the test's design, administration, scoring, reliability, validity, and fairness (Cardwell et al., 2024). EnglishScore uses mobile devices and reports proficiency through scales linked to the Common European Framework of Reference for Languages (CEFR); its official validity report describes the purpose, content, and reporting of the Core Skills Test (EnglishScore, 2021).

Differences in device, procedure, and task design may shape learners' experiences in different ways. User experience should be seen as part of the conditions under which scores are used, not as proof of their validity. Participants' responses can show whether procedures seem clear, fair, relevant, and accessible. Reviews of DETs, for instance, stress the need to carefully examine the relationship between test design, score interpretation, and intended use (Wagner, 202). In home testing, issues of proctoring, instructional language, privacy, devices, and connectivity also relate to fairness and access (Isbell & Kremmel, 202; Isbell et al., 2023).

This issue is relevant for PPG participants in Indonesia. Participants come from various geographic backgrounds, fields of study, digital experiences, and infrastructure conditions. In the EFL context, unfamiliarity with English-language instructions can mix with difficulty using the interface. As a result, participants may feel scores underrepresent their abilities not because there is psychometric evidence that the scores are invalid, but because the testing process feels unfamiliar, too fast, or technically disrupted. This conceptual distinction is important so that recommendations do not go beyond the available evidence. Research on computer-based assessments has addressed efficiency, adaptability, and validity, but direct comparisons of DET and EnglishScore user experiences in the context of Indonesian PPG are limited.

This study addresses this gap by integrating perception scores and open-ended comments. The goal is not to decide which test is psychometrically superior, but to map patterns of participant experiences and implementation conditions that institutions should consider. This study answers three questions: (1) how do participants rate access and ease of use, clarity of instructions, confidence in scores, usefulness of feedback, and technical readiness on DET and EnglishScore; (2) what barriers are most prominent in participants' comments; and (3) how do quantitative and qualitative findings reinforce or explain each other in assessing digital assessment experiences?

LITERATURE REVIEW

Computer-based language assessment and adaptivity

Computer-assisted language assessment involves using technology to present stimuli, receive responses, select items, process scores, and report results. In adaptive tests, item selection is influenced by ability estimates based on previous responses. This mechanism can improve measurement efficiency, but the quality of interpretation still depends on construct coverage, item bank quality, measurement model, and administration conditions (Douglas, 2007). So, the labels "digital" or "adaptive" do not automatically guarantee validity; evidence must be collected for specific score claims and uses. Chapelle and Douglas (2006) pointed out that evaluating computer-based tests requires considering the interaction between

technological characteristics and language proficiency. Interface, input type, response time, and technological literacy can affect performance. In digital-first tests, personalization and accessibility can reduce administrative barriers, but also require ongoing evaluation of design consequences (Maddox & Zumbo, 2021).

User experience, accessibility, and fairness

User experience in this study refers to participants' perceptions of the ease of accessing and using the platform, the clarity of instructions, technical fluency, and the usefulness of the results. These experiences matter because operational barriers can shift cognitive resources away from language tasks to managing devices or procedures. Structured instructions and clear displays are also linked to perceptions of test understandability and fairness, especially for participants new to digital formats. Remote administration increases options but raises questions about test environment equity. Isbell and Kremmel (202) suggest that home-based language tests should be evaluated with attention to devices, connectivity, security, and the consequences of high-stakes use. Isbell et al. (2023) further place remote proctoring within the perspective of fairness and justice: security procedures can burden groups of participants unequally, particularly when communication, technological access, or spatial conditions are not the same.

Perception of scores, assessment literacy, and feedback

Validity is not a fixed property of a test, but the strength of the arguments and evidence supporting the interpretation and use of the scores (Kane, 2013). Therefore, participants' responses to the statement "scores reflect ability" should be called perceptions of representativeness or trust in the scores, not psychometric validity. Perceptions are still valuable because they can show the acceptability, transparency, and consequences of the test experience, but they must be checked against evidence of reliability, internal structure, relationships with other measures, and impact of use. Assessment literacy helps participants understand the purpose of a test, the types of tasks, the criteria, and the meaning of the results reported. Yu and Xu (2024) showed that knowledge of assessment processes and criteria relates to participants' ability to critique assessments and use results for learning. Quick but overly general reports may have limited pedagogical value. In contrast, subskill scores, ability descriptors,

and follow-up instructions can help bridge the gap between measurement and learning.

Integrative framework of research

Based on the literature, the digital assessment experience is seen as the result of a three-layer interaction. The first layer is operational-technical: access, device compatibility, stability, audio, camera, and responsiveness. The second layer is procedural-cognitive: navigation, instructions, sample tasks, and time allocation. The third layer is interpretative-pedagogical: confidence in the representativeness of scores, clarity of reports, and usefulness of feedback. These three layers are used to organize analysis and interpret the convergence of quantitative and qualitative data.

RESEARCH METHOD

Research design

The study used a convergent mixed-methods design. Quantitative and qualitative data were collected through the same survey, analyzed separately, and then brought together during the interpretation stage. This design was chosen because average perceptions can show general patterns, while open-ended comments explain the specific forms of underlying difficulties (Creswell & Plano Clark, 2018). The study was exploratory and cross-sectional in nature.

Participants and context

Participants were PPG students in Indonesia who had used one of the assessment platforms. The DET survey received 197 visits, 150 sent responses, and 130 completed responses. The EnglishScore survey received 251 visits, 146 sent responses, and 134 completed responses. So, the analysis involved 264 completed responses. Partial responses were excluded. The sample was selected purposely based on direct experience using the platform. Because DET and EnglishScore participants were not the same group, comparisons were made at the level of descriptive patterns, not pairwise or causal comparisons.

Table 1. Survey response flow

Platform	Visit	Response sent	Complete responses are analyzed
DET	197	150	130
EnglishScore	251	146	134
Total	448	296	264

Instruments and procedures

The instrument consisted of 25 Likert-type statements on a 1-5 scale and one open-ended question asking for suggestions or impressions regarding difficult sections and challenges encountered during the test. Items covered accessibility and ease of use, clarity of instructions, perceived representativeness and reliability of scores, usefulness of feedback, and technical readiness. The survey was distributed online in April 2026 after participants completed the test. Instructions emphasized self-completion, honesty, and confidentiality of responses. The draft source stated that the instrument had undergone expert review and limited pilot testing. However, reliability statistics, validator details, and evidence of instrument validity were not available in the analyzed files. Therefore, this article does not make any claims of numerical reliability. Limitations in the instrument documentation are clearly noted and need to be addressed in replication research.

Data analysis

Quantitative analysis used item means and dimension means. The five dimensions each had five items, according to the instrument's order. The overall mean was calculated from the 25 item means. No significance tests, effect sizes, or confidence intervals were performed because individual data and standard deviations were unavailable. The 10-point satisfaction score listed in the initial draft was not used in the main synthesis because the conversion procedure was not documented. Open-ended comments were analyzed thematically through repeated readings, initial coding, code grouping, theme review, and theme naming. The procedure followed

thematic analysis principles, which require alignment between questions, analytical assumptions, and reporting (Braun & Clarke, 2022). Because the draft only provided relative occurrence rates and examples of comments, qualitative results are presented as thematic patterns, not absolute frequencies. Integration was achieved through a joint display that matched dimension means with comment themes.

RESULT AND DISCUSSION

Results

Quantitative patterns of user experience

EnglishScore achieved an overall mean of 4.34, while DET achieved 3.99. Across all five dimensions, EnglishScore's mean was consistently higher. The largest descriptive differences were in clarity of instructions ($\Delta = 0.42$), followed by confidence in the score ($\Delta = .36$), usefulness of feedback ($\Delta = .38$), technical readiness ($\Delta = .32$), and access and ease of use ($\Delta = .26$). This pattern suggests a more positive experience in the EnglishScore sample but does not show population differences or psychometric superiority of the platforms.

Table 2. Average user experience by dimension

Dimensions (5 items per dimension)	DET	EnglishScore	Descriptive differences
Access and ease of use	4.12	4.38	0.26
Clarity of instructions	3.89	4.31	0.42
Confidence in the score	3.92	4.28	0.36
Uses of feedback	3.96	4.34	0.38
Technical readiness	4.07	4.40	0.32
Overall average 25 grains	3.99	4.34	0.35

User constraint themes

Participant comments fell into five themes. First, unclear instructions and navigation, including confusion with speaking, listening, reading, and the

drag-and-drop mechanism. This theme was most prominent in the DET and also appeared in EnglishScore. Second, time pressure: some participants felt the time given was too short to understand the stimulus and respond. Third, audio and recording: participants were not always sure the audio was being recorded, had microphone difficulties, or judged the audio too quickly. Fourth, connection and stability issues: the platform stuttered or exited when the signal dropped. Fifth, limited feedback: participants wanted more detailed explanations of strengths and areas for improvement.

Table 3. Thematic summary of open comments

Theme	DET	EnglishScore	Illustration of comments from the draft
Instructions/navigation are not clear	Stand out	Appear	"The instructions for the speaking section are not clear."
Processing time	Appear	Appear	"The processing time felt fast."
Audio/recording	Appear	Appear	"Often don't know the recording has been recorded."
Stability/connection	Appear	Appear	"The platform stutters if the signal is bad."
Limited feedback	Limited	Limited	"Need more detailed feedback."

Integration of findings**Table 4.** Joint display of quantitative and qualitative integration

Quantitative findings	Qualitative findings	Meta-inference
Clarity of instructions is the lowest dimension of DET (3.89).	Confusion in task types and navigation is most prominent.	Procedural difficulties likely weigh on the testing experience beyond language ability.
Technical readiness is relatively high, especially EnglishScore (4.40).	Connection, audio, and recording complaints persist.	The majority of positive experiences do not negate the risks for participants with vulnerable technical conditions.
Confidence in scores is lower than technical access on both platforms.	Participants associated time, instructions, and recording with the representativeness of the results.	Score perceptions are formed from the administrative process, not just score reports.
The usefulness of feedback was rated positively, but not the highest.	Participants requested sub-skill details and follow-up.	Rapid reporting needs to be accompanied by actionable diagnostic information.

The joint display shows strong convergence: aspects with relatively lower averages receive direct explanation in the comments. At the same time, the comments reveal nuances not obvious from the averages, such as that technical issues can be rare but have significant consequences for individuals. There was not enough evidence to systematically assess divergence because the number of comments per theme was not documented.

DISCUSSION

User experience as a condition for score interpretation

Results show that participants can rate technical access positively while still having doubts about instructions and the representativeness of scores. This finding supports the idea that technology is not a neutral medium. If participants must guess the flow of a task, check recordings repeatedly, or adapt to unclear time limits, performance may be partly influenced by operational competence. Chapelle and Douglas (2006) and Douglas (2007) have stressed the need to assess the contribution of technology to the construct and condition validity of computer-based tests. The higher mean EnglishScore could be related to the mobile interface design, participants' familiarity with the device, the task format, or group characteristics. However, without random assignment, equal samples, and covariate data, the cause of the difference cannot be determined. The correct interpretation is that the EnglishScore user group in this study reported a more positive experience; not that EnglishScore is universally easier, more valid, or fairer than DET.

Instructions and time pressure

The largest gap in instruction clarity shows the importance of format orientation. Low scores on understanding each test section are consistent with comments about speaking, listening, reading, and drag-and-drop. Instructions serve as a bridge between the construct's purpose and participants' actions. When that bridge is unclear, errors may come from misinterpretation of procedures, not language limitations. This is relevant to assessment literacy: understanding the process and criteria helps participants develop suitable strategies and interpret the testing experience more accurately (Yu & Xu, 2024). Time pressure should be interpreted carefully. Time limits can be a legitimate part of the construct if processing speed is the target, but they can be an irrelevant source of variance if participants lose time learning the interface or handling the device. Organizers should distinguish between these two sources through formal, test-like exercises, clear time indicators, and opportunities to check device functionality before the test starts.

Infrastructure and fairness of distance assessment

Complaints about connection, microphone, audio, and recording suggest that high technical readiness scores should not be seen as an absence of risk. Averages can hide the experiences of disproportionately affected minorities. In remote testing, device or internet failures can interrupt the test, invalidate responses, or increase anxiety. These issues match concerns about access and equity in home-based testing (Isbell & Kremmel, 202; Isbell et al., 2023). For institutions in Indonesia, a fair response does not necessarily mean lowering security standards. Interventions can include compatibility tests, connection and device checklists, alternative test rooms for participants who need them, language-friendly support channels, and transparent recovery procedures. These policies reduce the chance that socio-technical conditions will affect outcomes.

Confidence in diagnostic scores and feedback

Statements about survey scores capture subjective beliefs, not validity. This distinction matters because validity needs a broader network of evidence (Kane, 2013). Still, participant perceptions should not be ignored. When participants see the process as unclear or responses may not be captured, confidence in the results can decrease, even if the scoring system is technically sound. Official DET and EnglishScore documentation provides information on score design and reporting (Cardwell et al., 2024; EnglishScore, 2021). For local use, institutions still need to help participants interpret scores. Results reports should include subskill mapping, ability descriptors, interpretation thresholds, and learning recommendations. This shifts scores from just numbers to information that can be used for teaching and learning.

Practical implications

Four implications can be drawn. First, organizers should provide a pre-test orientation with authentic task examples, navigation simulations, and timing guidelines. Second, microphone, camera, audio, operating system, and connection checks should be done before scoring sessions. Third, technical support and incident reporting procedures should be clear, prompt, and accessible. Fourth, score reports should be accompanied by diagnostic explanations and learning directions. These

recommendations apply to institutional implementation; changes to the platform's core design remain the responsibility of the test developer.

Limitations

The study has five limitations. First, the two platforms were assessed by different groups, which could have affected sample composition. Second, individual data, standard deviations, and demographic characteristics per group were not available in the draft, preventing inferential testing, effect size, and control analyses. Third, documentation of the validity and reliability of the survey instrument was incomplete. Fourth, the number of comments per theme and coder agreement were not available; therefore, the qualitative results are exploratory in nature. Fifth, perceived representativeness of scores cannot replace psychometric validity studies that compare scores to external measures or learning outcomes.

CONCLUSION

Based on 264 completed responses from PPG participants in Indonesia, EnglishScore achieved a higher average user experience rating than DET across the five dimensions analyzed. The most obvious gap was in the clarity of instructions, while comments highlighted difficulties with navigation, time pressure, audio or recording, and connectivity. Data integration showed that the quality of the digital assessment experience is shaped by technical readiness, procedural clarity, and the usability of result interpretation. The findings do not establish the psychometric superiority of either platform. Their main value lies in identifying implementation conditions that could be improved: pre-test orientation, task simulations, device checks, support for vulnerable connections, and more diagnostic score reporting. Further research should use the same participants on both platforms or use comparable group designs, retain item-level data, report instrument reliability, and link user perceptions to performance indicators and evidence of external validity.

REFERENCE

Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, 9(1), 3–26. <https://doi.org/10.1037/qup0000196>

- Cardwell, R., Naismith, B., LaFlair, G. T., & Nydick, S. W. (2024). Duolingo English Test: Technical manual. Duolingo, Inc.
https://duolingo-papers.s3.amazonaws.com/other/technical_manual.pdf
- Chapelle, C. A., & Douglas, D. (2006). Assessing language through computer technology. Cambridge University Press.
<https://doi.org/10.1017/CBO9780511733116>
- Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE.
- Douglas, D. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, 27, 115–132. <https://doi.org/10.1017/S0267190508070062>
- EnglishScore. (2021). Core Skills Test validity report: Relating language examinations to the CEFR—Describing your examination. EnglishScore Limited.
<https://www.englishscore.com/wp-content/uploads/2022/05/Validity-Study-2021-1.pdf>
- 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>
- Isbell, D. R., Kremmel, B., & Kim, J. (2023). Remote proctoring in language testing: Implications for fairness and justice. *Language Assessment Quarterly*, 20(4–5), 469–487. <https://doi.org/10.1080/15434303.2023.2288251>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Maddox, B., & Zumbo, B.D. (2021). Improving test validity and accessibility with digital-first assessments. Duolingo Research Report.
<https://duolingo-papers.s3.amazonaws.com/other/det-improving-test-validity.pdf>
- Wagner, E. (2020). Duolingo English Test, revised version July 2019. *Language Assessment Quarterly*, 17(3), 300–315. <https://doi.org/10.1080/15434303.2020.1771343>
- 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, Article 36. <https://doi.org/10.1186/s40468-024-00297-x>