Journal of Informatics and Vocational Education (JOIVE)
Vol.9, No.2, July 2026, pp. 67~83
ISSN: 2746-7813

Immersive TOEFL Preparation in the Metaverse: Usability and Navigability of a Roblox-Based Game Developed via GDLC

Rahmawati Salsabila[1], Rizki Hikmawan[2], Rhezwan Dhaifullah Romdhoni[3]
[1, 2, 3] Department of Education Systems and Information Technology, Universitas Pendidikan Indonesia

Article Info

Article history:

Received Apr 30, 2026
Revised May 3, 2026
Accepted Jun 1, 2026

Corresponding Author:

Rizki Hikmawan,
Universitas Pendidikan Indonesia, Indonesia
Email: hikmariz@upi.edu

ABSTRACT

Despite the growing adoption of game-based learning in language education, three-dimensional metaverse platforms for TOEFL preparation remain critically underexplored, leaving learners reliant on drill-based 2D media that lack immersion and sustained engagement. This study addresses that gap by developing a Roblox-based TOEFL learning game using the Game Development Life Cycle (GDLC) and evaluating user understanding of its game flow through graduated formative evaluation. An R&D design was employed, implementing six GDLC stages alongside Tessmer's formative evaluation: one-on-one (n=3), small group (n=5), and field testing (n=40). Data were gathered via observation, interview, and Likert-scale questionnaires, and analyzed descriptively. The game was realized as an area-based environment with three thematic zones and a multi-level navigation system. Evaluations showed progressive quality improvement: one-on-one (M=3.25), small group (M=4.00), and field testing (M=3.84, 96.3% positive response), indicating satisfactory usability and navigability. However, awareness of the Challenge Room and return-route comprehension remain areas requiring refinement. Theoretically, this study demonstrates that GDLC paired with formative evaluation provides a structured and iterative framework for validating metaverse-based educational games. Practically, the findings offer actionable design principles for educators and developers building immersive, game-based language learning environments within social 3D platforms.

Keywords: Game-based learning, Roblox, TOEFL, metaverse, GDLC

1. INTRODUCTION

English language skills are one of the important competencies for students in college, especially in academic and professional contexts. The Test of English as a Foreign Language (TOEFL) is a standard for measuring English language proficiency which is often used as a graduation requirement, prerequisites for final project exams, scholarship registration, and student exchange programs. However, the TOEFL preparation process often faces obstacles, especially in the listening part which requires repeated exposure, high focus, and the ability to capture verbal information in a limited amount of time. Commonly used practice media tends to be in the form of 2D quizzes or web-based practice questions, so the learning experience is more drill-oriented and less game-like interaction to maintain consistent learning engagement.

The game-based learning (GBL) approach has been reported to make the language learning experience more engaging because it encourages motivation and engagement to learn, as well as helps to reduce learning anxiety (Ahmed et al., 2022; Thurairasu, 2022). In line with that, metaverse-based learning environments have also been proven to be able to increase students' engagement, immersiveness, motivation, and practical skills, especially through gamification and collaborative virtual environments (Onu et al., 2024; Sripan & Jeerapattanatorn, 2025). In line with the development of 3D virtual environments, platforms like Roblox provide learning spaces that can be designed into immersive and social virtual worlds, allowing for a more

contextual world–room structure, scoring and reward systems, and feedback than 2D quizzes (Kye et al., 2021). Roblox in particular supports social and collaborative learning, and has been applied in a variety of higher education contexts such as virtual chemistry and STEM, with results showing increased interactivity and motivation compared to conventional media (Alhasan et al., 2023; Han et al., 2023; Onu et al., 2024). The study of Li & Yu (2023) It shows that a metaverse-based blended learning approach can improve academic success through immersive tasks, although the results are highly dependent on learners' digital literacy and self-regulation. This dependency on individual digital competence, however, is seldom controlled for in subsequent metaverse studies, raising questions about the generalizability of reported outcomes. Further, Park & Kim, (2022) discuss different types of gameful worlds in the metaverse, such as survival, maze, escape room, and interactive multiple choice, which are proven to increase learning motivation and support learning sustainability. In the context of TOEFL preparation, previous studies have utilized the game/gamification approach, but the use of 3D platforms such as Roblox as a TOEFL medium is still relatively rarely discussed specifically (Pratiwi & Waluyo, 2022).

A number of studies on language learning show that Roblox has been used in the context of EFL/ESL and tends to gain positive user perception, although there are still challenges, such as language barriers, during use (Sinar et al., 2023). Other research also confirms that the application of GBL to EFL can increase motivation and help reduce learning anxiety (Ahmed et al., 2022; Khoo et al., 2025). In terms of technology acceptance, a systematic study for 16 years was carried out by Chua & Yu, (2024) found that perceived usefulness and perceived ease of use are consistently the main determinants of acceptance or rejection of metaverse-based learning platforms. More specifically, Al-Adwan et al. (2023) in the extended TAM model found that perceived usefulness, personal innovation, and perceived enjoyment were the main driving factors for the intention to use metaverse-based learning platforms in universities, while the perception of cyber risk was a significant barrier. These findings are reinforced by Hwang et al. (2023) which shows that students' perceptions of 2D and 3D metaverse platforms are greatly influenced by social, interactive, and individual learning experiences. Taken together, these technology acceptance studies suggest that platform dimensionality alone does not predict user adoption; rather, social affordances and individual factors interact in ways that current TOEFL-specific GBL studies have yet to account for.

Despite this growing body of evidence, a critical gap remains: to the authors' knowledge, fewer than five peer-reviewed studies have specifically examined Roblox as a medium for TOEFL preparation, and none have applied a structured game development methodology combined with iterative formative evaluation to validate game flow comprehension before full deployment. Prior work either addresses GBL for general EFL contexts without platform specificity, or studies Roblox without a standardized assessment focus or systematic development rigor. This gap is significant because the absence of a validated design framework means that developers lack evidence-based guidance for building metaverse tools tailored to high-stakes language testing preparation.

Based on these needs, this study focuses on the development of Roblox-based TOEFL learning games using the Game Development Life Cycle (GDLC) and effective game design elements such as feedback, narrative, points, levels, and rewards proven to be key components in digital language learning games (Esteban, 2024; Govender & Arnedo-Moreno, 2021), so that it becomes a reference in the design process in this study. The formulation of this research problem is: (1) how the development process of Roblox-based TOEFL learning games uses the GDLC stage; and (2) how the results of the evaluation of user understanding of the TOEFL learning game flow were developed through formative evaluation consisting of one-on-one evaluation, small group evaluation, and field testing. The evaluation was carried out through three stages, namely one-on-one evaluation with 3 respondents for initial revision, small group evaluation for follow-up revision, and field testing for final validation, focusing on respondents' understanding of the flow, navigation, and mechanics of the game, not on the results of the TOEFL score obtained.

2. RESEARCH METHOD

2.1. RESEARCH DESIGN

This research is a research and development (R&D) project that focuses on the development of learning media in the form of Roblox-based TOEFL games. The product development process is carried out using the

Game Development Life Cycle (GDLC) method, which includes the initialization, pre-production, production, alpha-testing, beta-testing, and release stages. The product evaluation developed was carried out using formative evaluation to assess user understanding of the game flow through three stages, namely one-on-one evaluation, small group evaluation, and field testing.

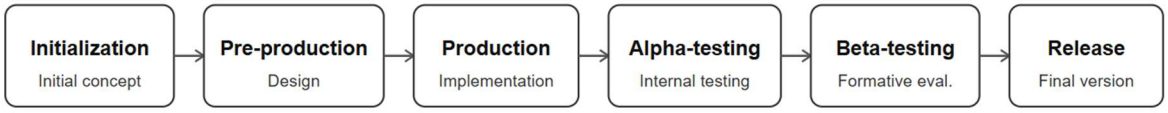

Figure 1. Research Design

2.2. RESEARCH SUBJECTS

The research subjects were students who became target users of Roblox-based TOEFL games and were selected *using purposive sampling techniques*. The research subjects were divided into three stages of formative evaluation: (1) One-on-One Evaluation Stage: 3 students for the initial trial; (2) Small Group Evaluation Stage: 5 students to test the results of the first stage of revision; and (3) Field Testing Stage: a larger number of students for final validation of the understanding of the game flow.

Figure 2. Research Subject

2.3. GAME DEVELOPMENT PROCEDURE (GDLC)

The selection of GDLC over other instructional design models such as ADDIE or Design-Based Research (DBR) is grounded in the nature of the artifact being developed. ADDIE is primarily designed for instructional content sequencing and lacks iterative playtesting mechanisms specific to interactive software, while DBR emphasizes theory generation through multiple macro-cycles that extend beyond the scope of a single development study. GDLC, by contrast, is purpose-built for game software development, offering structured phases from concept to release with built-in alpha and beta testing stages that align naturally with Tessmer's formative evaluation framework. The integration of GDLC with Tessmer's graduated evaluation one-on-one, small group, and field testing, allows iterative refinement of both technical functionality and user experience at each development phase, which is particularly suitable for validating game flow comprehension in a newly designed educational game environment.

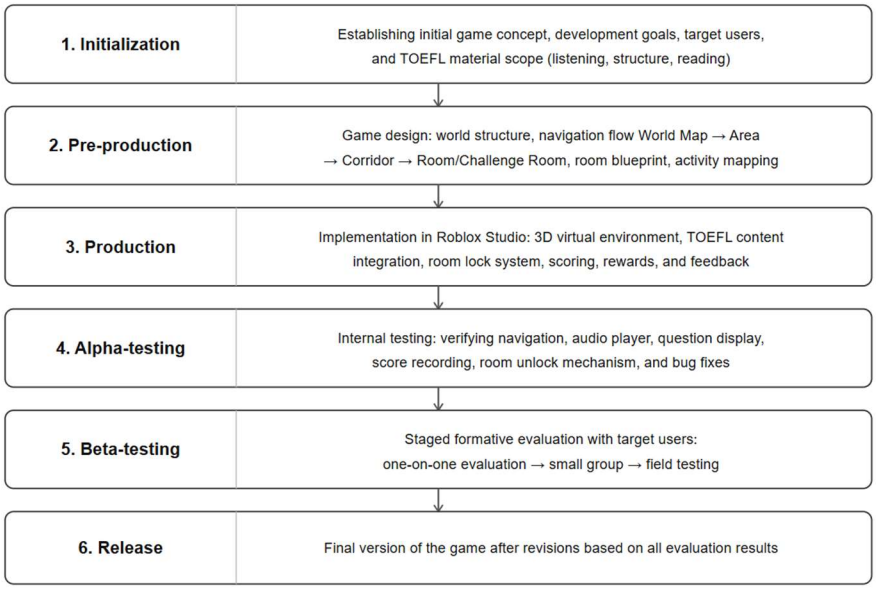

Figure 3. Game Development Procedure-GDLC

The game development process follows the GDLC stages which consist of six phases. At the *initialization stage, the* initial concept of the game, development objectives, target users, and scope of TOEFL (listening, structure, and reading) material are determined. The *pre-production* stage focuses on game design, including the game world structure, World Map navigation flow → Area → → Room/Challenge Room Corridor, level/room design, and learning activity mapping. The *production stage* is the implementation of the design into a game that can be played using Roblox Studio, including the creation of a 3D virtual environment, the integration of TOEFL materials, and the application of game mechanics (*room lock system,* score system, *rewards,* and feedback). The *alpha-testing* stage is carried out internally to ensure that all features run as planned. The *beta-testing stage* is carried out through a gradual formative evaluation of the target user. The *release stage* is the completion of the final version of the game after improvements have been made based on the results of the evaluation.

2.4. DATA ANALYSIS INSTRUMENTS AND TECHNIQUES

The research instruments consisted of: (1) observation guidelines for the one-on-one evaluation stage, (2) interview guidelines to explore the respondents' understanding in depth, and (3) a game flow understanding questionnaire that measured six indicators with a Likert scale of 5 points (1 = Strongly Disagree, 5 = Strongly Agree): (a) clarity of instructions, (b) ease of navigation, (c) understanding of objectives and flow, (d) clarity of game mechanics, (e) clarity of feedback, and (f) overall ease of use.

Qualitative data from observations and interviews (one-on-one stage) were analyzed by identifying problem patterns, classifying the types of difficulties, and formulating recommendations for improvement. Quantitative data from the questionnaire (small group stage and field testing) were analyzed descriptively by the following steps: (a) calculating the average score per item and per indicator; (b) calculate the percentage of positive responses; and (c) convert the average into interpretation categories. In One on One, Small Group and field testing, the instruments used in the session used a scale of 1–5, so the adaptive categories used were:

Table 1. Likert scale

Scale	Description
4,21–5,00	Excellent
3,41–4,20	Good
2,61–3,40	Pretty Good
1,81–2,60	Not Good
1,00–1,80	Not Good

3. RESULT AND ANALYSIS

3.1. Validity and Reliability of the Research Instrument

Prior to data collection, the questionnaire instrument was subjected to a content validity assessment conducted by one expert validator with expertise in educational technology and game-based learning design. The validator reviewed each instrument item for relevance, clarity, and alignment with the navigation construct being measured. Four of the six items were rated as valid without revision. Item 3, pertaining to pop-up utility in navigation decision-making, was flagged for minor wording revision to reduce ambiguity in how the action was described. Following the revision, the validator confirmed that all items were appropriate for administration. Given that only one validator was involved, formal Content Validity Ratio (CVR) computation was not applicable; validity assessment was therefore conducted through qualitative expert judgment in accordance with the single-expert review procedure. The results of the expert validation are summarized in Table 2.

Table 2. Content Validity Results — Expert Validation

Item	Navigation Dimension	Rating	Notes
Item 1	Ease of finding the area doors in the lobby	Valid	No revision required
Item 2	Clarity of the area name above the door	Valid with revision	Revision required
Item 3	Popup utility in decision-making	Valid with revision	Wording revised for clarity
Item 4	Popup text readability	Valid	No revision required
Item 5	Ease of locating teleport	Valid	No revision required
Item 6	Ease of finding Challenge Room	Valid	No revision required

Based on the expert validation results, all items were confirmed as content-valid following the revision process. Items 2 and 3 were the only items requiring adjustment, and the revised wording was reconfirmed by the validator before the instrument was administered to participants.

Following expert validation, an instrument readability check was conducted with two individuals who were not part of the main participant pool but shared similar characteristics with the target respondents university students with prior exposure to digital platforms. Both individuals were asked to read through all questionnaire items and indicate any items they found ambiguous or difficult to understand in terms of wording or response format. Based on their feedback, minor wording adjustments were made to two items to improve clarity, particularly in the phrasing of navigation-related action verbs. No structural changes to the instrument were required following the readability check.

Instrument reliability was then assessed using Cronbach's Alpha coefficient at each evaluation stage. The results are presented in Table 3.

Table 3. Cronbach's Alpha Reliability Coefficients by Evaluation Stage

Stage	n	Items (k)	Cronbach's α	Interpretation
One-on-one	3	4	0.952	Excellent*
Small Group	5	4	0.952	Excellent*
Field Testing	40	6	0.505	Moderate

At the one-on-one and small group stages, $\alpha = 0.952$ indicates excellent internal consistency among the four navigation items. However, these values should be interpreted cautiously given the very small sample size, as Cronbach's Alpha is sensitive to sample size and tends to be unstable when computed from fewer than ten respondents. At the field testing stage (n=40), $\alpha = 0.505$ reflects a moderate level of reliability. This moderate coefficient is attributable to the multidimensional nature of the navigation construct being measured the six items span distinct aspects of game flow including lobby orientation, visual labeling, pop-up utility, teleport accessibility, and Challenge Room awareness rather than representing a single unidimensional trait. To further examine item-level consistency at the field testing stage, item-total correlations were computed and are reported in Table 4.

Table 4. Item-Total Correlation Coefficients — Field Testing Stage (n=40)

Item	Navigation Dimension	r (Item-Total)
Item 1	Ease of finding area doors in lobby	0.073
Item 2	Clarity of area name above door	0.265
Item 3	Popup utility in decision-making	0.422
Item 4	Popup text readability	0.249
Item 5	Ease of locating teleport	0.378
Item 6	Ease of finding Challenge Room	0.206

Item-total correlations ranged from r = 0.073 (Item 1: lobby door ease) to r = 0.422 (Item 3: pop-up utility). Item 1 returned the lowest correlation, suggesting that ease of finding lobby doors is perceived as a relatively independent navigation dimension one that may be influenced more by prior Roblox experience than by the game's internal design cues. Item 3 returned the highest correlation, indicating that pop-up utility is the navigation dimension most representative of the overall game flow construct. The remaining items showed low-to-moderate correlations, consistent with the multidimensional interpretation of the instrument noted above.

Prior to conducting descriptive analysis, a Shapiro-Wilk normality test was applied to the total navigation score at the field testing stage to determine the appropriateness of the analytic approach. Results are presented in Table 5.

Table 5. Shapiro-Wilk Normality Test — Field Testing Total Navigation Score (n=40)

Test	Statistic	p-value	Conclusion
Shapiro-Wilk	W = 0.972	p = 0.411	Normal distribution (p > 0.05)

The Shapiro-Wilk test yielded W = 0.972, p = 0.411, indicating that the total navigation score distribution did not significantly deviate from normality. Given this result, descriptive statistics including means and percentage of positive responses were applied as the primary analytic method at the field testing stage. At the one-on-one and small group stages, normality testing was not conducted due to the very small sample sizes (n=3), which render such tests statistically uninformative; descriptive analysis was therefore applied directly at these stages as well.

Regarding methodological triangulation, data in this study were collected through three complementary sources: Likert-scale questionnaires (quantitative), structured observation records, and semi-structured interview responses (qualitative). This triangulation approach was applied across all evaluation stages to cross-validate findings and ensure that navigation usability was assessed from multiple perspectives. Qualitative data from observations and interview responses at the one-on-one and small group stages were independently coded by two researchers using a predefined coding scheme covering three categories: navigation difficulty patterns, feature awareness issues, and wayfinding behavior. Inter-rater agreement was assessed using Cohen's Kappa, with results presented in Table 6.

Table 6. Inter-Rater Reliability — Qualitative Coding (Cohen's Kappa)

Evaluation Stage	Coding Category	κ	Interpretation
One-on-one	Navigation difficulty patterns	0.81	Almost Perfect
Small Group	Awareness & wayfinding issues	0.76	Substantial
Overall		0.79	Substantial

The overall $\kappa = 0.79$ indicates substantial agreement between raters, suggesting that the qualitative coding process was sufficiently consistent and credible. Discrepancies between raters were resolved through discussion until consensus was reached before final coding was confirmed. Regarding data completeness, no questionnaire items were left unanswered at any evaluation stage; all submitted responses were complete and included in the analysis. No statistical outliers were detected at the field testing stage based on inspection of the score distribution, and all 40 responses were retained for analysis.

Having established the validity and reliability of the research instrument, the following sections present the results of each formative evaluation stage.

3.2. GDLC-Based Game Development Process

The development of Roblox-based TOEFL learning games is carried out through six stages of GDLC in a row and interdependent. At the initialization stage, it was determined that the game will be developed on the Roblox platform with the concept of an area-based learning environment that includes three TOEFL (Listening, Structure, and Reading) subject areas, targeting active students who are preparing for the TOEFL, with a learning approach based on game navigation and tiered challenges.

In the pre-production stage, a game structure was designed consisting of a World Map as a starting point for navigation, three learning areas, and within each area there is a corridor that connects three learning rooms (Room 1, 2, 3) and one Challenge Room as an evaluation per area. The navigation flow set is: World Map → Area → Corridor → Room/Challenge Room. Learning and evaluation activities per room are designed and outlined in the following blueprint.

Table 7. Learning Design of TOEFL Game-Based Modules (Listening, Structure, and Reading Areas)

Area	Room	Learning Objective	TOEFL Material	Game Activity	Assessment
Listening	Room 1	Understanding short conversations	Short Conversation	Listening to audio and answering multiple-choice questions	Correct score
	Room 2	Identifying detailed information	Long Conversation	Longer audio with tiered multiple-choice questions	Accumulated score
	Room 3	Understanding academic monologues	Talk/Monologue	Listening to monologues and answering inference-based questions	Accumulated score
	Challenge Room	Measuring comprehensive listening comprehension	Mixed listening	Randomized time-limited questions	Total score
Structure	Room 1	Understanding basic grammar	Basic Grammar	Completing sentence gaps	Correct score
	Room 2	Identifying structural errors	Error Recognition	Identifying incorrect parts of sentences	Accumulated score
	Room 3	Constructing sentence structure	Sentence Structure	Drag-and-drop or arrangement tasks	Accumulated score
	Challenge Room	Evaluating overall structure competence	Mixed structure	Randomized time-limited questions	Total score
Reading	Room 1	Identifying main ideas	Short Passage	Reading short texts and answering main idea questions	Correct score
	Room 2	Understanding vocabulary in context	Vocabulary in Context	Determining word meanings in context	Accumulated score
	Room 3	Analyzing longer texts	Long Passage	Text analysis and inference questions	Accumulated score
	Challenge Room	Evaluating overall reading competence	Mixed reading	Randomized time-limited questions	Total score

At the production stage, the design is implemented using Roblox Studio. The 3D virtual environment is built with a separate area system, inter-room corridors, a room lock system, a real-time scoring mechanism, and feedback (true/false) answers. The integration of TOEFL material is done through interactive objects in the game (computers, chairs, screens). The alpha-testing stage is carried out internally to verify the navigation function, audio player, question presentation, score recording, and room opening mechanism. All identified bugs were fixed before entering the beta-testing stage, which was carried out through a gradual formative evaluation. After going through all stages of evaluation and revision, the product enters the release stage as the final version that has met the functionality and usability aspects. At this stage, the system is declared stable and suitable for use as a game-based learning medium in the context of TOEFL learning.

3.3. Format Evalution

3.3.1. One-on-One Evalutation

One-on-one evaluation was carried out by involving three students (S-01, S-02, S-03) who met the inclusion criteria. Quantitative data was collected through four items of navigation statements using a 5-point Likert scale, while qualitative data was obtained through direct observation and in-depth interview sessions after the play session. Table 4 presents the characteristics of the respondents.

Table 8. Karakteristik Responden One-on-One Evaluation

Code	TOEFL Experience	Frequency of Gaming	Areas of Navigation Difficulty
S-01	Familiar with the TOEFL format	Frequent	Enter the Challenge Room area
S-02	Has/are preparing	Sometimes	Enter the area from the World Map
S-03	Has/are preparing	Frequent	Enter the area from the World Map

Table 3 presents the characteristics of respondents involved in the one-on-one evaluation stage. The results indicate that all participants have prior exposure to TOEFL, either being familiar with the test format or currently preparing for it. In terms of gaming experience, most respondents frequently engage in digital games, which supports their ability to interact with the game-based learning environment.

However, several navigation difficulties were identified, particularly when entering specific areas such as the Challenge Room and accessing areas from the World Map. These findings provide initial insights into potential usability issues that need further evaluation.

Table 9. Results of the One-on-One Evaluation Phase Navigation Comprehension Questionnaire (Scale 1–5)

No.	Navigation Statement Item	S-01	S-02	S-03	Mean	Category
1	Ease of moving between areas of the World Map	3	4	3	3,33	Good
2	Ease of finding Rooms 1, 2, 3, and Challenge Rooms in each area	3	4	3	3,33	Good
3	Clarity of in-game signs and directions	3	4	3	3,33	Good
4	Not confused about the order of the room after completing one room	3	4	2	3,00	Pretty Good
Overall Average					**3,25**	Pretty Good

Table 4 shows the results of the navigation comprehension questionnaire during the one-on-one evaluation stage. Three out of four navigation items achieved a mean score of 3.33, categorized as “Good,” indicating that respondents generally found the navigation features easy to understand. However, the fourth item, related to the clarity of room sequence after completing a task, obtained a lower mean score of 3.00, categorized as “Fair.”

Overall, the average score of 3.25 falls into the “Fair” category, suggesting that while the navigation system is generally acceptable, certain aspects still require improvement. Specifically, only 66.7% of respondents gave positive responses (score ≥ 3) to the fourth item, as one respondent (S-03) reported confusion regarding the sequence of navigation between rooms.

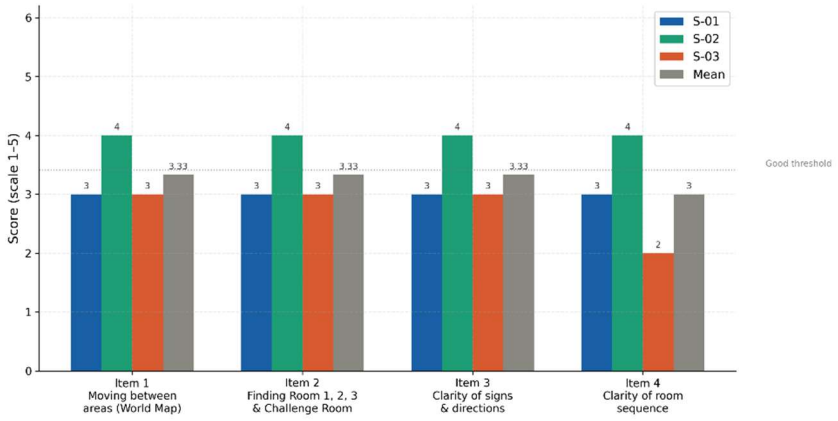

Figure 4. Navigation Scores-One-on-One Evaluation

Figure 4 illustrates that three out of the four navigation items achieved mean scores of 3.33, indicating a generally positive evaluation of navigation usability. In contrast, the fourth item, related to the clarity of room

sequence after completing a task, obtained a lower mean score of 3.00, suggesting relatively greater difficulty in this aspect.

Overall, the average score reached 3.25, which falls into the "Fair" category. Furthermore, the percentage of positive responses (score ≥ 3) for items 1–3 reached 100%, while item 4 only achieved 66.7%. This result is influenced by one respondent (S-03) who assigned a score of 2, indicating confusion regarding the sequence of navigation between rooms.

Qualitative data analysis identified three major patterns of navigation problems, as shown in Table 5.

Table 10. Problem Navigation Patterns of One-on-One Evaluations and Improvement Recommendations

No.	Problem Patterns	Description of Findings	Recommended Improvements
1	Confusion of the entry mechanism of the area from the World Map	S-02 and S-03 have trouble finding and activating the entrance of the area from the World Map (not understanding the interaction buttons)	Added visual cues (arrows/glow) and pop-up instructions to area doors on the World Map
2	The order of the room is not always clear	S-03 gives a score of 2 for clarity of room order; S-01 accidentally entered the Challenge Room while exploring the area near the elevator	Strengthening of room order labels (Room 1 → 2 → 3 → Challenge Room) with more striking visual indicators
3	Absence of contextual clues in the room	S-01 suggests that each room have a pop-up message of instructions so that players immediately understand the purpose and how to play	Addition of instructional hint popups at the beginning of each room and Challenge Room

Based on the identified problem pattern, three main revisions were made before entering the next stage: (1) the addition of a visual cue to the area door on the World Map; (2) strengthening of the room massage label; and (3) the addition of a hint popup at the beginning of each room.

3.3.2. Small Group

Small group evaluation was conducted with three new respondents using a more comprehensive instrument covering seven dimensions of navigation on a scale of 1–5, to test the effectiveness of revisions made after one-on-one evaluation. Table 6 presents the results of this stage questionnaire.

Table 11. Results of the Small Group Evaluation Navigation Evaluation Stage (Scale 1–5, n=5)

No.	Navigation Aspect	Red	Category
1	Ease of finding area doors in the lobby	4,00	Good
2	Clarity of the name of the area above the door	4,67	Excellent
3	Uses of room name popups	3,67	Good
4	Readability of popup text (size, color, position)	4,00	Good
5	Ease of locating teleports within the area	3,67	Good
6	Ease of finding Challenge Rooms	4,00	Good
7	Ease of return routes between areas	4,00	Good
Overall Average		**4,00**	**Good**

Based on Table 6, the overall average score of navigation reached 4.00 out of 5.00, which falls into the "Good" category. This result indicates that the navigation system has improved and is generally well understood by users following the revisions made after the one-on-one evaluation.

Among the evaluated aspects, the clarity of area names displayed above the doors achieved the highest mean score (4.67), categorized as "Very Good." This suggests that the enhancement of area labeling was effective in improving user navigation. Meanwhile, the usefulness of room name pop-ups and the ease of locating teleport features obtained relatively lower mean scores (3.67), indicating that these elements still require further refinement.

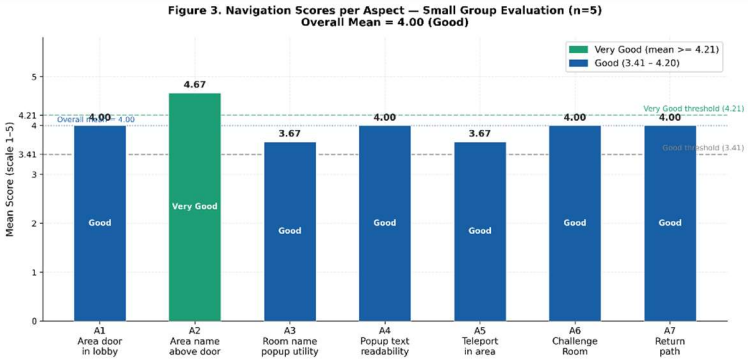

Figure 5. Navigation Scores-Small Group Evaluation

Figure 5 further illustrates the distribution of navigation scores across all evaluated aspects. The visualization confirms that most navigation elements received consistently high ratings, indicating improved usability after the revision stage. However, slightly lower scores in the pop-up and teleport features highlight specific areas where navigation clarity can still be enhanced.

Table 12. Qualitative analysis of small group sessions

Code	Frequency of Gaming
S-01, S-05	Two respondents needed exploration time before finding the teleport, each stating "go first" or "look around" before succeeding
S-01, S-03, S-04	Room name popups aren't always immediately noticed, with one respondent stating "not paying attention" and the other two "only waking up after a while
S-02	respondents only found out about the existence of the Challenge Room after exploring the area, not from the initial information of the game
S-02	Doubts on the return lane

Compared to the one-on-one stage, there was an indicated improvement in the area label clarity aspect (4.67 vs. 3.33 on different scales), which confirmed the effectiveness of post-revision area marker reinforcement. However, teleport awareness and Challenge Room awareness emerged as new problem points that needed to be addressed before field testing. Qualitative data from the small group stage provided more granular insight: two respondents (S-01, S-05) required extended exploration time before locating the teleport, using language suggesting a trial-and-error strategy rather than intentional navigation ("go first," "look around"). Three respondents (S-01, S-03, S-04) did not notice room name pop-ups until mid-session, with one explicitly stating "not paying attention" — suggesting that the pop-up placement and timing did not align with users' natural visual scanning patterns. One respondent (S-02) was entirely unaware of the Challenge Room's existence until late exploration, indicating that the feature lacked sufficient introductory signposting. These findings confirm that awareness-level issues require proactive onboarding design, not merely visual enhancement.

3.3.3. Field Testing

Field testing was conducted with 40 students who interacted with the game across several scheduled sessions. It should be noted that the available data represent a partial dataset from the overall target of the field testing; therefore, the findings reflect preliminary patterns that require confirmation through complete data collection.

Table 13. First Area Visit Pattern by Field Testing Respondents (n=40)

First Area Visited	Frequency (n)	Percentage (%)
Listening	24	60,0%
Structure	8	20,0%
Reading	8	20,0%
Total	40	100,0%

Based on Table 8, 60% of respondents selected the Listening area as the first area to visit, followed by Structure and Reading, each at 20%. The most common sequence of area visits was Listening → Structure

→ Reading (L→S→R), observed in 42.5% of respondents. Other sequences, such as Structure → Reading → Listening, appeared less frequently.

This pattern suggests that respondents tend to follow the conventional TOEFL section order when exploring the game environment, indicating that the game design aligns well with users' prior expectations.

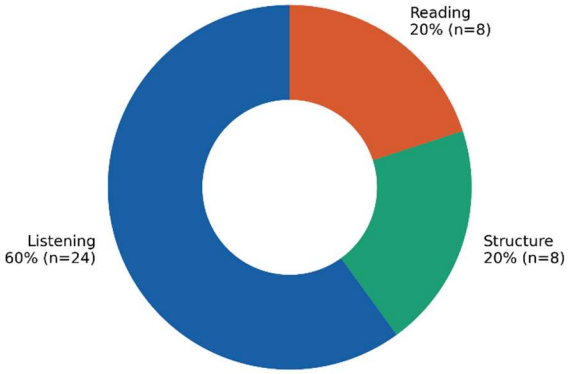

Figure 6. First Area Visited by Responded

Figure 6 visually confirms that the Listening area was the most frequently selected entry point. The dominance of this area highlights users' preference to begin with familiar TOEFL sections, reinforcing the alignment between game structure and conventional test sequencing.

Table 14. Results of Evaluation of the Navigation Dimensions of the Field Testing Stage (Scale 1–5, n=40)

No.	Navigation Dimensions	Mean	Positive (%)	Category
1	Ease of finding area doors in the lobby	3,80	95,0%	Good
2	Clarity of the name of the area above the door	4,08	100,0%	Good
3	The use of the room name popup in decision-making	3,70	87,5%	Good
4	Readability of popup text (size, color, position)	3,80	100,0%	Good
5	Ease of locating teleports within the area	3,85	97,5%	Good
6	Ease of finding Challenge Rooms	3,83	97,5%	Good
Overall Average		**3,84**	**96,3%**	**Good**

Based on Table 9, all six navigation dimensions fall within the "Good" category, with an overall mean score of 3.84 out of 5.00 and a positive response rate of 96.3%. This indicates that the navigation system is generally well accepted by users in the field testing stage.

The highest score was obtained for the clarity of area names displayed above the doors (mean = 4.08; 100% positive), confirming the effectiveness of visual labeling. In contrast, the usefulness of room name pop-ups received the lowest mean score (3.70; 87.5% positive), suggesting that this feature still requires further improvement.

Additionally, the minimum scores observed in dimensions 1, 3, 5, and 6 reached 2 (on a 5-point scale), indicating that a small proportion of respondents still experienced difficulties in certain navigation aspects.

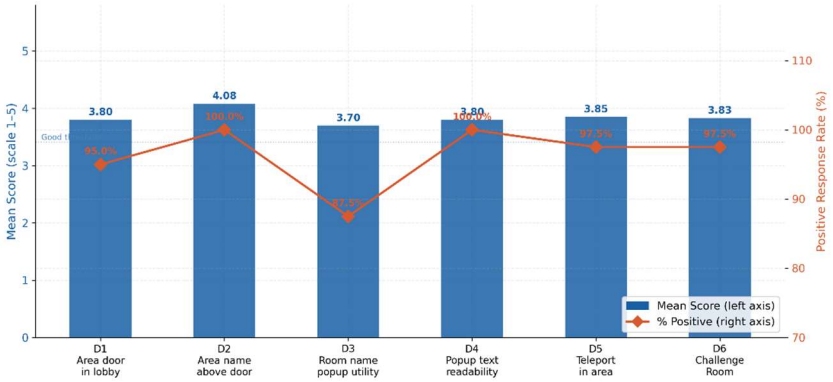

Figure 7. Navigation Scores-Small Group.

Figure 7 illustrates the distribution of navigation scores across all evaluated dimensions. The visualization confirms that most navigation aspects received consistently high ratings, indicating stable usability performance at the field testing stage. However, slight variations in lower-scoring dimensions highlight areas that may benefit from further refinement.

Table 15. Categorical Findings of the Field Testing Stage (n=40)

Aspects	Yes / Confused	No / Not Confused	Others
Entering the wrong room in the lobby	10 (25,0%)	30 (75,0%)	—
Confusion on the way back between areas	17 (42,5%)	20 (50,0%)	3 (7,5%)
Not Knowing the Challenge Room from the Beginning	4 (10,0%)*	25 (62,5%)	11 (27,5%)**

** Not knowing the existence of the Challenge Room; ** Only found out after exploration*

Based on Table 10, the majority of respondents did not experience errors in initial navigation, as 75.0% did not enter incorrect rooms in the lobby area. This finding indicates that the visual cues and initial navigation structure are generally clear and effective for most users.

Regarding return navigation between areas, 50.0% of respondents reported no confusion, while 42.5% experienced some difficulty. This suggests that improvements are still needed in spatial orientation and directional guidance to enhance clarity without compromising usability.

In terms of awareness of the Challenge Room, only 10.0% of respondents were unaware of its existence from the beginning. Meanwhile, 62.5% were able to locate it directly, and 27.5% discovered it through exploration. These results indicate that the feature is generally accessible, and the game environment supports independent discovery.

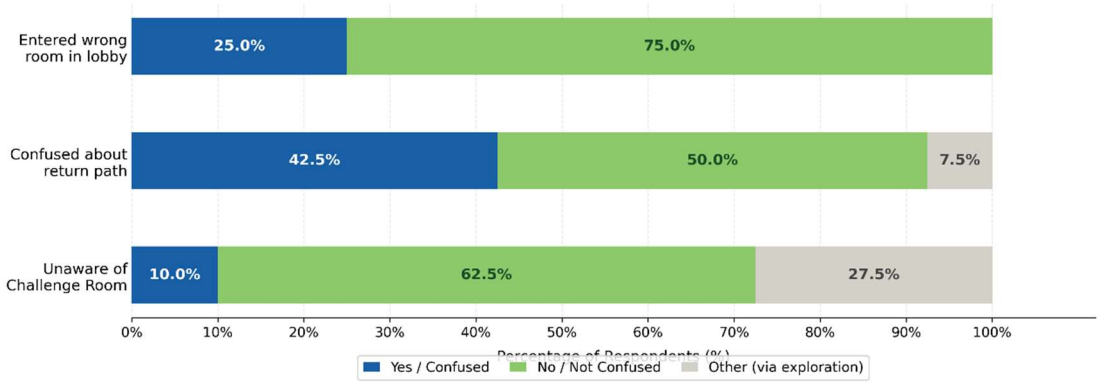

Figure 8. Categorical Findings

Figure 8 further illustrates the distribution of categorical responses across navigation aspects. The visualization reinforces that most users were able to navigate the system effectively, with only a minority experiencing confusion in specific areas. These findings suggest that the system functions well overall, with minor usability improvements needed, particularly in navigation flow and initial user guidance.

3.4. ROBLOX-Based TOEFL Game Development via GDLC

The game development process using GDLC resulted in an area-based learning environment structured on the Roblox platform, answering the first research question. This success is in line with the findings (Kye et al., 2021) which found that Roblox inherently supports a virtual world-based learning design, including spatial navigation structures, material integration, and interactive mechanisms. These findings are also reinforced by Han et al. (2023) In a systematic review of the use of Roblox in learning found that the platform inherently supports a social and collaborative learning environment, as well as being able to facilitate learners' cognitive and noncognitive skills through a structured virtual world design. Furthermore, Han et al. (2023) affirm that Roblox's high appeal to young users as well as its support for VR-based environments makes it a potential platform for formal learning, including language test preparation such as TOEFL, as long as it is supported by a robust instructional design. This confirms that the choice of the Roblox platform in this study is not just a technical consideration, but is based on empirical evidence about the platform's suitability for structured learning needs. The GDLC, with its six successive stages, provides a systematic development framework from conceptualization to release so that every aspect of the game's design and functionality can be tested and verified in stages before being presented to users.

The concept of an area-based learning environment is designed with three material areas (Listening, Structure, Reading), each containing three gradual rooms and one Challenge Room that corresponds to the gameful world-type taxonomy in the metaverse proposed by Park & Kim, (2022). The multi-choice world type combined with the time-limited evaluation elements of the Challenge Room reflects a combination of maze (navigational exploration) and multi-choice (choice-based evaluation) mechanics that, according to Park & Kim, (2022) Especially effective in supporting the motivation of continuous learning. The design of the game elements in this study, particularly the use of feedback, graded levels, and scoring mechanisms in the Challenge Room, was in line with the analysis Govender & Arnedo-Moreno, (2021) which identifies that feedback, points, levels, and clear goals are the most consistent design elements found in effective digital game-based language learning. In line with that, Esteban, (2024) in a systematic review of digital game-based language learning (DGBLL) it was confirmed that these elements not only increase learning engagement, but also support the acquisition of vocabulary and language skills in a contextual manner. Further, the integration of TOEFL material into Roblox's 3D environment replicates the model that was successfully done (Alhasan et al., 2023) in the context of virtual chemistry, the presentation of academic content through 3D games has been shown to increase interactivity and motivation compared to conventional media. Thus, the integration of these elements into the Roblox environment developed in this study represents the application of the empirical evidence-based DGBLL principles.

3.5. The Effectiveness of Gradual Formative Evaluation in Identifying Usability Issues

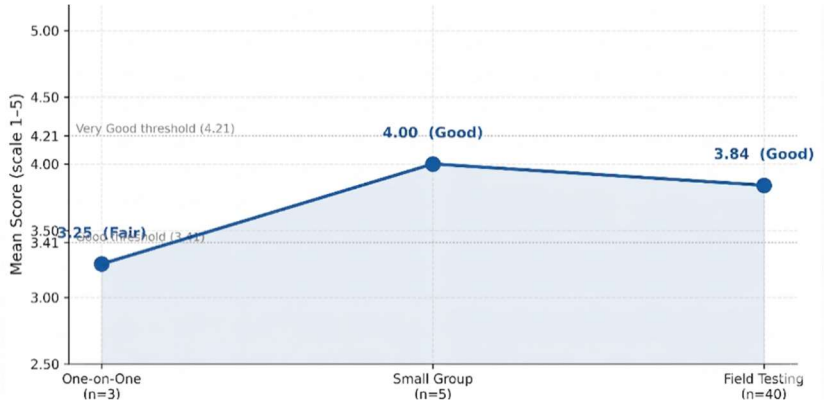

Figure 9. Mean Score Progression Across Formative Evaluation Stages

The results of the three-stage formative evaluation answered the second research question and showed the methodological consistency of the Tessmer model in identifying and progressively improving usability problems. In the one-on-one evaluation stage, three patterns of fundamental navigation problems were identified: (1) the unintuitive door interaction mechanism of the World Map, (2) the unclear sequence of the room, and (3) the absence of contextual clues in the room. Qualitatively, observation notes from the one-on-one stage revealed that all three respondents paused for more than 15 seconds when first encountering the lobby door, and two of the three verbally expressed uncertainty ("I don't know where to go"). Interview responses further indicated that the absence of directional text or arrow cues near the door was the primary source of disorientation. These behavioral and verbal indicators collectively pointed to a wayfinding design gap rather than a content comprehension issue, which directly informed the revision of area door labels prior to the small group stage. An average score of 3.25/4.00 (Good) at this stage indicates that a basic understanding of navigation has been established, but still needs to be refined at some critical points.

The effectiveness of the post-one-on-one evaluation revision was reflected in the results of the small group evaluation, which showed an improvement in the area name clarity aspect (mean = 4.67/5.00; Very Good), aspects that are directly addressed through the revision of the strengthening of area door labels. An overall average of 4.00/5.00 in the small group confirmed that the improvements made were generally effective. However, the emergence of new problems, especially in teleport awareness and Challenge Room awareness, suggests that a gradual evaluation is indeed needed to uncover layers of problems that are not visible in small samples. This pattern is consistent with the logic of Tessmer's formative evaluation, where each stage is designed to uncover a different problem contextually.

At the field testing stage with n=40, an overall average of 3.84/5.00 (Good category) and a positive response rate of 96.3% indicated that the game was rated navigable by most respondents on a broader scale. It is worth noting that the overall mean score decreased from 4.00 (small group) to 3.84 (field testing). This decline is interpreted not as a regression in product quality, but as a natural scaling effect: the small group (n=5) consisted of purposively selected participants with relatively homogeneous digital literacy profiles, whereas the field testing sample (n=40) introduced greater variation in prior experience with 3D virtual environments. This pattern is consistent with the observation of Hwang et al. (2023) that individual differences in social and interactive experience significantly influence perceived navigability in metaverse platforms. The score reduction therefore, reflects increased sample heterogeneity rather than decreased usability and underscores the importance of designing for a wider range of user competencies in future iterations.

These results are consistent with the findings Li & Yu, (2023) which shows that metaverse-based learning environments tend to produce positive judgments when participants have adequate exploration and self-regulation skills. The entire dimension is in the Good category. This demonstrates that the integration of GDLC with Tessmer's formative evaluation produces a measurable and staged improvement trajectory from M=3.25 (one-on-one) to M=4.00 (small group) to M=3.84 (field testing), constituting an empirical contribution to the application of Tessmer's model in game-based learning contexts. Theoretically, this study extends Tessmer's framework by demonstrating its applicability not only to instructional material evaluation but also to game flow and navigability assessment in 3D virtual environments. Within the GBL literature, the findings confirm that structured iterative evaluation is a necessary condition for achieving acceptable usability in educational game design, adding nuance to the claim that GBL inherently produces engagement without specifying the design conditions required. From a TAM perspective, the high positive response rate (96.3%) at field testing provides quantitative evidence that perceived ease of use, operationalized here as navigability, is achievable in metaverse-based learning tools when iterative user-centered design is applied.

3.6. The Dynamics of Challenge Room Awareness and Return Paths as Critical Design Points

Findings at the field testing stage show that most users are able to interact with the system effectively, including in accessing key features such as the Challenge Room. Although there were 10.0% of respondents who did not know the existence of the feature from the beginning, the majority of users still managed to find it, either directly (62.5%) or through the exploration process (27.5%). This indicates that the navigation structure and design of the game environment have been able to facilitate the discovery process independently, so that the main features remain accessible to most users without any significant obstacles.

Within the framework of the Technology Acceptance Model (TAM), put forward by Al-Adwan et al. (2023), Perceived ease of use is an important factor in determining the acceptance of metaverse-based learning platforms. When a portion of users (27.5%) recognize the existence of a Challenge Room through exploration, it indicates that an understanding of the feature is not fully established in the early stages of interaction, but is still evolving through the use of the system directly. This condition leads to the need to strengthen the aspect of feature recognition at the beginning of use. In line with that, Chua & Yu, (2024) emphasizing that clarity of initial orientation is an important element in increasing the acceptance of metaverse-based learning technologies. Thus, these findings are more accurately understood as the need for onboarding optimization and visual emphasis, rather than as a fundamental weakness of the system.

In the aspect of return routes between areas, although 42.5% of respondents experienced confusion, half of the respondents (50.0%) did not experience difficulties in navigating again. These findings suggest that the designed navigation mechanism is understandable to most users, although there are still variations in the spatial orientation experience. This indicates that the system has provided a functional navigation base, but it still needs to be refined to improve the consistency of the user experience. This is in line with Hwang et al. (2023), which states that the perception of ease of navigation in the metaverse environment is influenced by the individual characteristics of users. Overall, the findings on the Challenge Room and return pathways aspects reflect the need for iterative enhancement in advanced development, without compromising the fact that the system has achieved a good level of usability and navigability in the context of game-based learning.

3.7. Game Design Implications for Game-Based Learning TOEFL Learning

The pattern of area visits in field testing, where 60% of respondents chose the Listening area first with the most popular order L→S→R, indicates the tendency of respondents to follow the conventional order of the TOEFL test. This shows that respondents bring their TOEFL model mentality into game navigation, which is a positive indicator that the game structure is successfully communicating with the user's prior knowledge. These findings are relevant to the argument Kye et al., (2021) Roblox facilitates social and cognitive learning through a game structure that can be tailored to specific academic contexts.

From a broader GBL perspective, the fact that 96.3% of respondents responded positively to game navigation supports the findings Ahmed et al., (2022); Khoo et al. (2025) which states that the GBL approach consistently results in positive assessments of media and learning motivation. This high positive assessment is also consistent with the findings Thurairasu, (2022) which confirms that the gamification approach in language learning consistently produces a positive perception of the media and encourages the intrinsic motivation of students to continue to engage. In addition, Esteban, (2024) emphasizes that the success of digital game-based language learning does not depend solely on the linguistic content presented, but is heavily influenced by the quality of the gaming experience itself, including clarity of purpose, ease of navigation, and relevance of the game's context to the learning objectives. The fact that the structure of the TOEFL is successfully communicated through the design of the game, as reflected in the L→S→R visit pattern that follows the conventional test sequence, suggests that the game developed has successfully integrated the principles of the DGBLL into the context of formal test preparation. Based on these findings, several specific and actionable implications are offered for educators and game developers working in similar contexts. First, educators intending to use this game as a supplementary TOEFL preparation tool should conduct a brief onboarding session (10–15 minutes) prior to independent play, specifically orienting students to the Challenge Room location and the return-route mechanism between areas, the two navigation aspects that showed the highest confusion rates. Second, developers designing metaverse-based educational games should incorporate mandatory introductory waypoints or guided first-play sequences that surface all major features before open exploration begins, rather than relying on ambient discovery. Third, given that 27.5% of respondents discovered the Challenge Room through exploration rather than direct instruction, future iterations should add a persistent mini-map or HUD (heads-up display) indicator that marks key feature locations throughout gameplay. Fourth, for institutions considering adoption, it is recommended that the game be piloted with a small cohort (5–10 students) and navigation comprehension be assessed via a brief post-play checklist before full-class deployment, to account for individual differences in 3D environment literacy. More specifically, the Challenge Room elements, designed as time-limited evaluations, reflect the mechanism, according to Park & Kim, (2022) Especially effective at maintaining learning engagement in a metaverse environment. However,

the effectiveness of this mechanism can only be optimal if the respondents manage to find a gap that is currently still identified as a critical point.

The findings on perceived enjoyment and user acceptance, implicitly reflected in the positive navigation score, are in line with the TAM model developed Al-Adwan et al. (2023), which places perceived enjoyment as a strong predictor of intention to use metaverse platforms in college. A high positive response (96.3%) indicates that the game manages to deliver a generally enjoyable and non-cognitively taxing experience, a condition that, according to UN et al., (2024) is an important prerequisite for the sustainability of using metaverse platforms as learning media.

4. LIMITATIONS

Several limitations of this study warrant explicit acknowledgment. First, the field testing dataset represents a partial sample from the overall target population; data were collected from 40 respondents across scheduled sessions, and the findings should therefore be treated as preliminary patterns pending confirmation through complete data collection. This partial dataset constitutes a serious constraint on the generalizability of the field testing results, and conclusions drawn from this stage should be interpreted with appropriate caution. Second, the study focused exclusively on game flow and navigation comprehension as the primary evaluation construct; no measurement of actual TOEFL learning outcomes or score improvement was conducted, limiting the ability to draw conclusions about the game's instructional effectiveness. Third, expert validation was conducted by a single validator, which limits the robustness of the content validity assessment. Fourth, inter-rater reliability for qualitative coding was established between two researchers within the same research team, which may introduce shared perspective bias.

5. FUTURE RESEARCH

Future research directions are proposed in direct response to the limitations identified above. First, to address the partial dataset limitation, subsequent studies should ensure complete field testing data collection with the full intended sample before drawing generalizability conclusions; a minimum sample of 80–100 respondents is recommended to support more robust inferential analysis. Second, given that this study was limited to navigation and game flow evaluation, future research should incorporate a pre-post quasi-experimental design to measure the actual effect of the Roblox-based TOEFL game on listening, structure, and reading scores, using a validated TOEFL proficiency instrument as the outcome measure. Third, to strengthen content validity, future iterations of this instrument should involve a panel of at least three expert validators to enable formal CVR and CVI computation following Lawshe's (1975) procedure. Fourth, to address the onboarding weaknesses identified at the field testing stage, specifically low initial awareness of the Challenge Room (10%) and return-route confusion (42.5%), development should implement and empirically test a structured first-play tutorial sequence, measuring whether guided onboarding significantly reduces navigation errors compared to the current open-exploration design. Fifth, future studies should examine the moderating role of individual differences, particularly prior Roblox experience and digital literacy, on navigation performance and game acceptance, using a survey instrument such as the Technology Acceptance Model (TAM) scale administered alongside the navigation evaluation.

6. CONCLUSION

This research successfully developed a Roblox-based TOEFL learning game through the Game Development Life Cycle (GDLC) approach, producing a structured, interactive, and immersive area-based learning environment consisting of three thematic areas and a multi-level navigation system. The iterative development process, supported by Tessmer's graduated formative evaluation, proved effective in systematically identifying and improving usability aspects across three stages: one-on-one (M=3.25), small group (M=4.00), and field testing (M=3.84; 96.3% positive response). These results confirm that the integration of GDLC with formative evaluation provides a viable and structured framework for developing navigable metaverse-based educational games. Furthermore, the game structure that aligns with the conventional TOEFL

section sequence demonstrates that the resulting design is not only technically functional but also pedagogically relevant in supporting language test preparation.

REFERENCES

Ahmed, A. A. A., Ampry, E. S., Komariah, A., Hassan, I., Thahir, I., Hussein Ali, M., Fawzi Faisal, A., & Zafarani, P. (2022). Investigating the Effect of Using Game-Based Learning on EFL Learners' Motivation and Anxiety. *Education Research International*, *2022*. https://doi.org/10.1155/2022/6503139

Al-Adwan, A. S., Li, N., Al-Adwan, A., Abbasi, G. A., Albelbisi, N. A., & Habibi, A. (2023). "Extending the Technology Acceptance Model (TAM) to Predict University Students' Intentions to Use Metaverse-Based Learning Platforms". *Education and Information Technologies*, *28*(11), 15381–15413. https://doi.org/10.1007/s10639-023-11816-3

Alhasan, K., Alhasan, K., & Al Hashimi, S. (2023). Roblox in Higher Education. *International Journal of Emerging Technologies in Learning (IJET)*, *18*(19), 32–46. https://doi.org/10.3991/ijet.v18i19.43133

Chua, H. W., & Yu, Z. (2024). A systematic literature review of the acceptability of the use of Metaverse in education over 16 years. *Journal of Computers in Education*, *11*(2), 615–665. https://doi.org/10.1007/s40692-023-00273-z

Esteban, A. J. (2024). Theories, Principles, and Game Elements that Support Digital Game-Based Language Learning (DGBLL): A Systematic Review. *International Journal of Learning, Teaching and Educational Research*, *23*(3), 1–22. https://doi.org/10.26803/ijlter.23.3.1

Govender, T., & Arnedo-Moreno, J. (2021). An analysis of game design elements used in digital game-based language learning. *Sustainability (Switzerland)*, *13*(12). https://doi.org/10.3390/su13126679

Han, J., Liu, G., & Gao, Y. (2023). Learners in the Metaverse: A Systematic Review on the Use of Roblox in Learning. *Education Sciences*, *13*(3). https://doi.org/10.3390/educsci13030296

Hwang, Y., Shin, D., & Lee, H. (2023). Students' perception on immersive learning through 2D and 3D metaverse platforms. *Educational Technology Research and Development*, *71*(4), 1687–1708. https://doi.org/10.1007/s11423-023-10238-9

Khoo, Y. Y., Ramdan, M. R., Abdullah, N. L., Abd Aziz, N. A., & Mahjom, N. (2025). The Impacts of Game-based Learning on Thinking and Learning in Higher Education Context: A Scoping Review. *International Journal of Education in Mathematics, Science and Technology*, *13*(3), 623–637. https://doi.org/10.46328/ijemst.4776

Kye, B., Han, N., Kim, E., Park, Y., & Jo, S. (2021). Educational applications of metaverse: Possibilities and limitations. *Journal of Educational Evaluation for Health Professions*, *18*. https://doi.org/10.3352/jeehp.2021.18.32

Li, M., & Yu, Z. (2023). A systematic review on the metaverse-based blended English learning. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1087508

Onu, P., Pradhan, A., & Mbohwa, C. (2024). Potential to use metaverse for future teaching and learning. *Education and Information Technologies*, *29*(7), 8893–8924. https://doi.org/10.1007/s10639-023-12167-9

Park, S., & Kim, S. (2022). Identifying World Types to Deliver Gameful Experiences for Sustainable Learning in the Metaverse. *Sustainability (Switzerland)*, *14*(3). https://doi.org/10.3390/su14031361

Pratiwi, D. I., & Waluyo, B. (2022). Integrating Task and Game-Based Learning Into an Online Toefl Preparatory Course During the Covid-19 Outbreak At Two Indonesian Higher Education Institutions. *Malaysian Journal of Learning and Instruction*, *19*(2), 37–67. https://doi.org/10.32890/mjli2022.19.2.2

Sinar, T. S., Budiman, M. A., Ganie, R., & Rosa, R. N. (2023). Students' Perceptions of Using Roblox in Multimodal Literacy Practices in Teaching and Learning English. *World Journal of English Language*, *13*(7), 146–153. https://doi.org/10.5430/wjel.v13n7p146

Sripan, T., & Jeerapattanatorn, P. (2025). Metaverse-based learning: A comprehensive review of current trends, challenges, and future implications. *Contemporary Educational Technology*, *17*(3). https://doi.org/10.30935/cedtech/16434

Thurairasu, V. (2022). Gamification-Based Learning as The Future of Language Learning: An Overview. *European Journal of Humanities and Social Sciences,* *2*(6), 62–69. https://doi.org/10.24018/ejsocial.2022.2.6.353