OPEN ACCESS

Technology in Language Teaching & Learning
ISSN 2652-1687
https://www.castledown.com/journals/tltl/

Castledown

Technology in Language Teaching & Learning, 8, 103687 (2026)
https://doi.org/10.29140/tltl.2026.103687

Play to Learn: A Meta-analysis on the Impact of Game-based Learning on the Language Learning Outcomes of Young and Middle-aged Adults

CHRISTIAN AMOR T. TUSO[a] (Corresponding author)
MARY ALYSSA A. ALCONABA[b]
JULIUS B. OPRIASA[c]
ANGELICA T. FERNANDEZ[d]
GEORGINA ISABEL D. BALAURO[e]

[a]*Research Development and Innovation Center, Our Lady of Fatima University, Laguna, Philippines*
cat.tuso24@gmail.com

[b]*Senior High School Department, Our Lady of Fatima University, Laguna, Philippines*
maalconaba@fatima.edu.ph

[c]*College of Business and Accountancy, Our Lady of Fatima University, Laguna, Philippines*
jbopriasa@fatima.edu.ph

[d]*College of Physical Therapy, Our Lady of Fatima University, Laguna, Philippines*
altiongco@fatima.edu.ph

[e]*College of Arts and Sciences, Our Lady of Fatima University, Laguna, Philippines*
gdbalauro@fatima.edu.ph

Abstract

Background: Game-Based Learning (GBL) is an emerging instructional methodology that uses game mechanics to enhance individual learning experiences. This meta-analysis examined the differential impact of game-based learning compared to non-game-based learning (NGBL) methods among young and middle-aged adults. **Methods:** Comparative studies exploring the difference between NGBL and GBL among young and middle-aged adults were obtained from different databases. The literature search yielded a total of 2,695 citations. Study selection and evaluation reduced this number to six studies included in the meta-analysis, which were then analyzed using Review Manager. Risk of bias assessment (ROBINS-I and Rob 2.0) and sensitivity analyses were performed to further assess the consistency of the pooled estimate. **Results and Discussion:** Meta-analysis using a random-effects model indicated that GBL was associated with better language learning outcomes compared to NGBL methods in studies included in the meta-analysis. A statistically significant large effect size was observed (SMD = 0.92, 95% CI [0.58, 1.25], $p < 0.0001$). Moderate to substantial heterogeneity ($I^2 = 55\%$), and borderline non-significant variability (Cochran's $Q = 11.02$, $p = 0.05$) were observed among the studies included in the meta-analysis. Risk of bias assessment and sensitivity analyses showed low to moderate concerns, suggesting that the pooled estimate was directionally consistent across tested assumptions. **Conclusion:** The findings of the study suggest that GBL may be associated with better language learning outcomes compared with NGBL methods and may serve as a useful supplementary instructional approach for reinforcing targeted language skills.

Keywords: Game-Based Learning (GBL); meta-analysis; adults; random-effects model; language.

Introduction

Digital innovations and developments in teaching strategies have influenced educators to explore new ways of enhancing learner's education experiences. One such development is Game-Based Learning (GBL), an instructional approach that leverages the interactive and engaging nature of games to enhance learning (Wu, 2018; Roohani & Vincheh, 2021). This contrasts with Non-Game-Based Learning (NGBL), which includes traditional or conventional methods such as lectures and teacher-led discussions, as well as alternative approaches such as self-directed or student-centered learning that utilize existing media and instructional materials to facilitate learning (Jitpaisarnwattana & Saville, 2025; Topal, 2025). While gamification simply involves incorporating game mechanics (e.g. leaderboards, achievements, and badges) to enhance existing lecture materials, GBL utilizes the game itself as the primary medium for delivering subject content (Al-Azawi et al., 2016; Zafar et al., 2014). Previous systematic reviews examining the effectiveness of GBL have highlighted its potential to promote learning engagement and increase motivation across a wide range of subject areas (Yasir et al., 2021; Pan et al., 2021; Sousa et al., 2023).

Despite its potential, there is no certain proof regarding the effectiveness of GBL compared to NGBL methods. Its reported efficacy is often dependent on different factors such as game design, type of learner, and the subject being discussed (Yasir et al., 2021). Although previous studies support the effectiveness of GBL in improving the language ability of young and middle-aged adults (Chen et al., 2023; Zafar et al., 2014), challenges and limitations are also present, such as technical glitches and the alignment of the game's characteristics with the specific learning outcomes that are needed (Cabrera-Solano, 2022; Ogawa, 2025). This is evident especially in analog or non-digital GBL, which was found to be effective in the acquisition of soft skills (i.e. communication and collaboration) among students but has limited impact on the improvement of their content or hard skills (Sousa et al., 2023). Digital GBL approaches may have more expansive pedagogical uses but are frequently hindered by the resource-intensive nature of game design, which requires substantial time, technical expertise, and investment (Pan et al., 2021).

Several meta-analyses have already investigated the impact of game-based learning on cognitive outcomes (Dixon et al., 2022; Barz et al., 2023; Liao et al., 2010). However, some of these reviews took into account only pre- and post-test scores in intervention groups, making direct comparisons with NGBL methods impossible. Previous study also included a varying range of participants in terms of their age, which could influence the result since individuals from different age groups have different learning motivations and performance (Caldwell-Harris & MacWhinney, 2023). Sigmundsson et al. (2022) reported that passion and grit vary across age groups. Individuals aged 19 years and below tend to exhibit a stronger growth mindset compared to older adults aged 70–79, while demonstrating slightly lower levels of grit than those in the young and middle-aged adult groups. Notably, both Dixon et al. (2022) and Barz et al. (2023) provided similar medium effect sizes for GBL interventions (0.50 and 0.54, respectively), indicating overall consistency in the effectiveness of GBL in enhancing cognitive functions. Dixon et al. (2022) emphasized improvements in language learning specifically, while Barz et al. (2023) generalized their results to the improvement of cognitive, metacognitive, and affective-motivational outcomes. Liao et al. (2010), on the other hand, attempted to compare GBL with traditional methods in various fields of knowledge but provided little information on the included studies, so their conclusions remained vague and inapplicable.

To address these gaps and further examine the pedagogical potential of GBL, the present review aimed to assess the overall association of game-based learning, compared to non-game-based learning methods, with language-related outcomes among young and middle-aged adults. Specifically, it examined differences between the two learning approaches across various language-related measures through meta-analysis and sought to provide evidence on whether GBL may serve as a potentially useful supplementary instructional approach for language learning within the target population.

Methods

Search Strategy

This study was guided by the registered OSF protocol (osf.io/4mp5t) and presented according to the preferred reporting items for systematic review according to Page et al. (2020). The following databases were utilized for the acquisition of literature: Scopus, Education Resources Information Center (ERIC), PubMed, OpenAlex, JSTOR, and Web of Science. Similar terminologies were used for all databases; however, different search strings were utilized depending on the database's guidelines. The terms "gamification" and "gamified learning" are included since some articles classify purely game-based learning as gamified due to the presence of game mechanics. This was later on checked and verified during the screening of the articles, whether the articles are concerned with gamified or game-based learning. For Scopus/ERIC/OpenAlex: (gamification OR gamify OR "game-based" OR "game inspired" OR "play*") AND ("young adult*" OR adolescen* OR juvenile OR teenag* OR "middle adult*" OR "mid-life" OR "middle-aged") AND (languag* OR "language proficiency" OR "language ability") AND (education OR learn* OR teach*). For JSTOR: (gamification OR gamify OR "game-based") AND ("young adult*" OR "middle-aged" OR "middle adult*") AND (languag* OR "language proficiency" OR "language ability") AND (education OR learn* OR teach*). For PubMed: (gamification[Title/Abstract] OR gamify[Title/Abstract] OR "game-based"[Title/Abstract] OR "game inspired"[Title/Abstract] OR "play*" [Title/Abstract]) AND ("young adult*"[Title/Abstract] OR adolescen*[Title/Abstract] OR juvenile[Title/Abstract] OR teenag* [Title/Abstract] OR "middle adult*" [Title/Abstract] OR "mid-life" [Title/Abstract] OR "middle-aged" [Title/Abstract]) AND (languag*[Title/Abstract] OR "language proficiency"[Title/Abstract] OR "language ability" [Title/Abstract]) AND (education[Title/Abstract] OR learn*[Title/Abstract] OR teach* [Title/Abstract]). For Web of Science: (TI= (gamification OR gamify OR "game-based" OR "game inspired" OR "play*") AND ("young adult*" OR adolescen* OR juvenile OR teenag* OR "middle adult*" OR "mid-life" OR "middle-aged") AND (languag* OR "language proficiency" OR "language ability") AND (education OR learn* OR teach*)) OR (AB= (gamification OR gamify OR "game-based" OR "game inspired" OR "play*") AND ("young adult*" OR adolescen* OR juvenile OR teenag* OR "middle adult*" OR "mid-life" OR "middle-aged") AND (languag* OR "language proficiency" OR "language ability") AND (education OR learn* OR teach*)) OR (KP= (gamification OR gamify OR "game-based" OR "game inspired" OR "play*") AND ("young adult*" OR adolescen* OR juvenile OR teenag* OR "middle adult*" OR "mid-life" OR "middle-aged") AND (languag* OR "language ability"))). The search strings were utilized on April 17, 2025, and additional citations through hand search were done on April 22, 2025. Hand searching was done by utilizing the following terminologies: Game-based learning, Traditional Teaching, Alternative Learning, Language proficiency, Young and Middle adults.

Study Selection and Quality Assessment

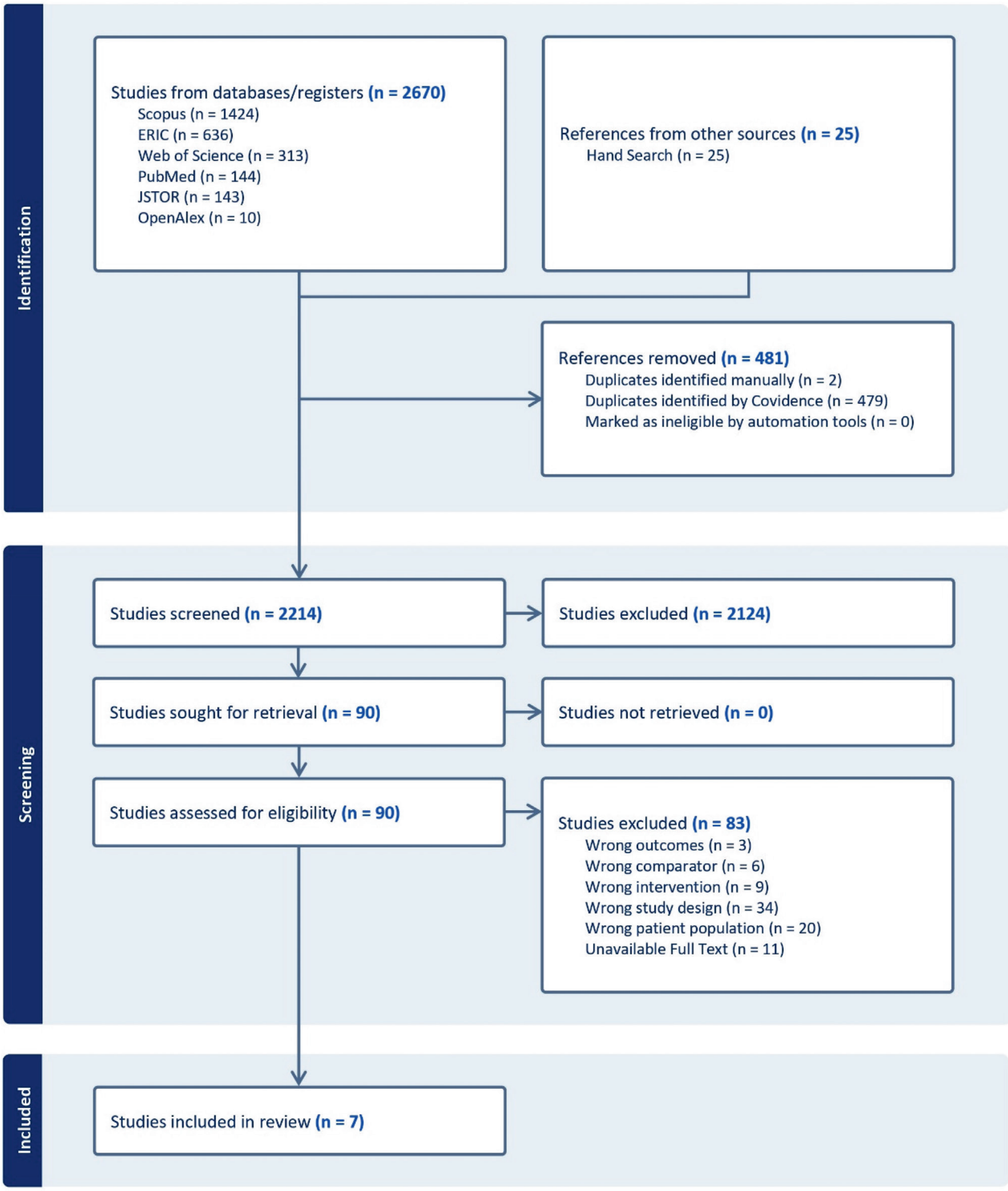

Figure 1 *PRISMA Flow Diagram of the Literature Review and Selection of Studies*

The review was conducted following the PRISMA flow diagram presented in Figure 1. The articles included were restricted to those written in English, had full-text availability, and followed the following criteria (PICO). Participants (P) of their study should be those who belong to young and middle-aged adults with an age range of 18 to 59 years old (Erikson, 1994 as cited in Infurna et al., 2020). Studies will still be included if they involve younger or older individuals, provided that the mean age of the sample falls within the defined range of young to middle-aged adults. Additionally, studies that do not specify an exact age range but describe their participants as belonging to a certain population group classified as young to middle-aged adults will also be considered. The intervention (I) must include the use of Game-based Learning, while studies that solely implement gamification techniques, such as the addition of game mechanics (leaderboards, achievements, points, etc.), were excluded. To be included, a study had to employ a comparative design involving another learning condition (C); therefore, studies that assessed the impact or effectiveness of Game-based Learning without a control or comparison group were excluded. These comparison conditions included traditional or teacher-led classroom instruction, as well as self-directed or student-centered methods. To more accurately reflect this variability, the present review collectively refers to these conditions as non-game-based learning (NGBL). The studies included should have a Control Group (CG) who undergone NGBL such as teacher-led discussions or self-learning approaches, and an Experimental Group (EG) that received GBL, either digital or analog. The outcomes (O) should include both pre-test and post-test scores of participants, regardless of the assessment tool used. If possible, reported inferential statistics (e.g. t-test) should be included in the reporting of baseline scores. Studies that did not apply statistical analysis in the participants' scores will be analyzed first using an independent t-test using the indicated pre- and post-test scores and standard deviations. Furthermore, the inclusion of both pre- and post-tests allows for the utilization of change scores as primary effect measure for the meta-analysis. This is to better account for the baseline differences between experimental and control groups, particularly in studies with non-randomized designs (Deeks et al., 2022).

Table 1 *Category and Criteria for the Mixed Methods Appraisal Tool (MMAT) by Hong et al. (2018)*

Category of study designs	Quality criteria
2: Quantitative randomized controlled trials	2.1: Is randomization appropriately performed?
	2.2: Are the groups comparable at baseline?
	2.3: Are there complete outcome data?
	2.4: Are outcome assessors blinded to the intervention provided?
	2.5: Did the participants adhere to the assigned intervention?
3: Quantitative non-randomized	3.1: Are the participants representative of the target population?
	3.2: Are measurements appropriate regarding both the outcome and intervention (or exposure)?
	3.3: Are there complete outcome data?
	3.4: Are the confounders accounted for in the design and analysis?
	3.5: During the study period, is the intervention administered (or exposure occurred) as intended?

Each included article was evaluated using the Mixed Methods Appraisal Tool (MMAT) by Hong et al. (2018) to assess their quality. Different domains of the study were assessed and graded from 0% to 100%. MMAT was deemed as the appropriate tool for quality assessment since both Randomized Controlled Trial (RCT) and Non-Randomized Controlled Trial (NRCT) were included in the review.

Hong et al. (2018) mentioned that the appraisal tool has no specific characterization of interpretation, and it could depend on the author on how they interpret their results. For this study, studies that fall within 0–20% grade are considered as low quality, 40–80% as medium quality, and 100% as high quality. The criteria of the appraisal tool were summarized following the representations shown in Table 1.

Data Extraction

All citations retrieved from the databases utilized in the study were uploaded to the Covidence Systematic Review Management Software, where duplicates were automatically removed by the system. Two (2) votes were required for an article to proceed, following the PRISMA model presented in Figure 1. Title and abstract were screened by three independent reviewers (MA, JO, and CA), and conflicts were resolved through a consensus between the three. This was done through a discussion based on the inclusion/exclusion criteria of the study. The same process was done for the Full-text screening of the articles uploaded that passed the title and abstract screening. A spreadsheet was used for an organized and collaborative extraction of the information from the included studies. Literature information, Methodology, Population and sample size, and results of the studies that passed the full-text screening were recorded and assessed. In the case where a study tested 2 different parameters, both results were considered and were treated as a different set of experimental data. Lastly, when immediate and delayed post tests were obtained, only the immediate results were included in the data extraction to ensure consistency within the included studies.

Included studies that reported two learning outcomes derived from the same set of samples were further processed by treating each outcome as statistically dependent. Outcomes that measure the same language domain were combined into a single study-level effect size to avoid unit-of-analysis errors as recommended by the Cochrane Handbook (Higgins et al., 2024). In case correlations between outcomes were not explored, a default value for correlation (r = 0.5) was used in the following equation:

$$SD_{combined} = \frac{\sqrt{SD_1^2 + SD_2^2 + 2r(SD_1 - SD_2)}}{2} \tag{1}$$

Where the standard deviation of the sample groups (CG and EG) is represented as SD, and r indicates the correlation between the reported measures. The combined mean values of the learning measures were averaged assuming that both measures were derived from the same sample size and pertained to the same language domain.

$$Change\ score_{mean} = Post_{mean} - Pre_{mean} \tag{2}$$

$$SD_{change} = \sqrt{SD_{pre}^2 + SD_{post}^2 - 2r(SD_{pre} * SD_{post})} \tag{3}$$

When change scores and their corresponding standard deviations were not reported, the following equations (2 and 3) were used to derive them. Pre- and post-tests language outcomes from the included studies were converted into change scores using Equation 2. A similar correlation coefficient of r = 0.5, consistent with that used for the combined SD, was assumed to calculate for the standard deviation of the change scores. Sensitivity analyses were also conducted for both the combined and change score standard deviations using r = 0.3 and r = 0.7 to ensure that results were not affected by the assumed correlations.

For multi-arm studies, only intervention arms that are relevant to the review question and meeting the inclusion criteria were retained. Non-eligible arms were excluded to avoid inappropriate comparisons and double counting, in line with the recommendations of the Cochrane Handbook for handling multiple intervention groups (Higgins et al., 2024). It should be noted that the selection of relevant intervention groups was made independently of statistical significance. Although not all intervention groups were included in the meta-analysis, all intervention groups in multi-arm studies should still be reported. However, only those included in the meta-analysis need to be described in detail (Higgins et al., 2024).

Risk of Bias Assessment

ROBINS-I by Sterne et al. (2016) was used to assess the potential bias of NRCT studies across seven domains: (1) confounding, (2) classification of interventions, (3) selection of participants, (4) deviations from intended interventions, (5) missing data, (6) measurement of outcomes, and (7) selection of reported results. RoB 2.0, by Sterne et al. (2019), was used to assess the potential bias in RCTs within the five domains: (1) the randomization process, (2) deviations from intended interventions, (3) missing outcome data, (4) measurement of outcomes, and (5) selection of reported results. Both tools were developed by the Cochrane Collaboration and are available via their website (https://www.riskofbias.info/welcome). Two independent raters, Ma and CA, assessed the risk of bias within the included studies individually, and any disagreement was resolved through discussions. Robvis (Risk of bias visualization) was then used to present the assessment results of each included study (McGuinness & Higgins, 2020).

Meta-analysis

The fixed-effect model was not applicable due to the variability in the sample size and research tools of the included studies; hence, a random-effects meta-analysis was employed (Jain et al., 2019; Schünemann et al., 2023). RevMan (Version 5.4.1; Cochrane Collaboration, 2020) was used to conduct the meta-analysis, applying the inverse variance method and analyzing the heterogeneity between the included studies through the I^2 statistics and Cochrane's Q test. The null hypothesis in Cochran's Q suggests that there is no considerable heterogeneity between the included studies, and a p-value of 0.05 was used due to the lack of power for the test (Chang et al., 2022). Interpretation of I^2 values was based on the Cochrane Handbook guidelines indicating a minimal heterogeneity for values ranging from 0 to 40%, moderate heterogeneity for 30 to 60%, a substantial heterogeneity for 50 to 90%, and a considerable heterogeneity for I^2 values ranging from 75 to 100% (Deeks et al., 2022). Chang et al. (2022) further suggested that I^2 values less than 40% may suggest unimportant heterogeneity.

The standardized mean difference (SMD) based on change scores, reported in RevMan as Hedges' g, for each included study was calculated using inverse-variance random-effects models and was used to express the effect size of the overall analysis. It was selected due to its ability to account for differences in standard deviations among the results of the study and to correct biases that may be incurred due to small sample sizes (Brydges, 2019). Effect sizes were then expressed as SMD (Hedges' g) using the following equation:

$$Hedges'g = \frac{\bar{X}1 - \bar{X}2}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}} \left(\frac{N-3}{N-2.25}\right)\left(\frac{\sqrt{N-2}}{N}\right) \tag{4}$$

Where $\bar{X}$ represents the mean of the sample groups (CG and EG), SD denotes the standard deviation, and n indicates their corresponding sample sizes. The second part of the equation, which includes N (total sample size), represents the bias correction factor applied when the sample size is small, particularly when $N < 50$. The interpretation of SMD (Hedges' g) followed Schünemann et al. (2023), where values of 0.20 to 0.49 indicate a small effect, 0.50 to 0.79 indicate a moderate effect, and values ≥ 0.80 indicate a large effect.

$$95\%\ \text{Confidence Interval}\ (CI) = \text{Hedges}'\ g_{overall} \pm 1.96\ (SE_m) \tag{5}$$

The analysis included the reporting of totals and subtotals, the 95% confidence intervals (CI) for each study, and the overall pooled SMD (Hedges' g). Prediction intervals were then calculated using Equation 6 to estimate the range within which the effect size of a future study would be expected to fall if randomly added to the meta-analysis (Spineli & Pandis, 2020).

$$\text{Prediction Interval} = \text{Hedges}'\ g_{overall} \pm t_{\alpha,k-2}\sqrt{\tau^2 + SE_m} \tag{6}$$

Where τ^2 represents the between-study variance in a random-effects meta-analysis, SE_m represents the standard error of the pooled effect, while $t_{\alpha,k-2}$ represents the critical value at k-2 degrees of freedom with k representing the number of the included studies. Since the standard error of the pooled effect is not included in the results projected by RevMan, it is calculated from the confidence interval as shown in equation 5.

The robustness of the meta-analysis was assessed through sensitivity analysis via the one-study-removal method, adapted from the methodology of Ru et al. (2025). This was done through the sequential removal of individual studies to assess the changes in the overall effect size as well as the significance of the meta-analysis (Deeks et al., 2022). Furthermore, subgroup-removal was also done in order to assess the effect of different study designs. A minimal to no deviation from the original overall effect size indicates that no individual study overinfluenced the result of the meta-analysis (Aung et al., 2026). Subgroup comparison utilized chi-square test to determine possible differences between the included subgroups. Additionally, studies with unresolved inconsistencies between reported values and inferential results were subjected to additional verification through full-text re-examination. When discrepancies cannot be resolved with confidence, such studies are excluded from the primary meta-analysis and examined separately in sensitivity analysis.

Results

A total of 2,670 citations were retrieved from Scopus, Web of Science, PubMed, ERIC, JSTOR, and OpenAlex, with an additional 25 citations obtained through manual searching. Of the 2695 total citations uploaded, 479 duplicates were detected by Covidence, and two (2) additional duplicates were manually identified. This resulted in 2,214 studies remaining for title and abstract screening, which resulted in the further exclusion of 2,124 citations due to irrelevance. Subsequently, 90 full-text articles were assessed, of which 83 were excluded. The remaining seven (7) studies were initially included prior to the meta-analysis, as illustrated in Figure 1.

Article Distribution

Table 2 presents the publication characteristics of the studies who met the eligibility criteria (n = 7). An equal proportion of studies involved multiple and single authorship (n = 2 each; 28.57%). Most arti-

Table 2 *Studies Meeting the Review Eligibility Criteria (n = 7)*

Author(s), year	Article title	Source title	Country (region)
Zafar et al., 2014	Game-based learning with native language hint and their effects on student academic performance in a Saudi Arabia community college	Journal of Computers in Education	Saudi Arabia (EM)
Chen et al., 2023	A game-based learning system based on octalysis gamification framework to promote employees' Japanese learning	Computers & Education	Taiwan (WP)
Cabrera-Solano, 2022	Game-Based Learning in Higher Education: The Pedagogical Effect of Genially Games in English as a Foreign Language Instruction	International Journal of Educational Methodology	South America (A)
Wu, 2018	Improving the effectiveness of English vocabulary review by integrating ARCS with mobile game-based learning	Journal of Computer Assisted Learning	Taiwan (WP)
Tang & Taguchi, 2021	Digital Game-Based Learning of Formulaic Expressions in Second Language Chinese	The Modern Language Journal	United States (A)
Roohani & Vincheh, 2021	Effect of game-based, social media, and classroom-based instruction on the learning of phrasal verbs	Computer Assisted Language Learning	Iran (EM)
Alhebshi & Gamlo, 2022	The Effects of Mobile Game-Based Learning on Saudi EFL Foundation Year Students' Vocabulary Acquisition	Arab World English Journal	Saudi Arabia (EM)

Note: EM = Eastern Mediterranean, WP = Western Pacific, A = American.

cles (n = 3; 42.86%) originated from countries in the Eastern Mediterranean region and were written through partner authorship (n = 3; 42.86%). Notably, most of the articles were published in computer science and education journals and appeared within a 10-year period, from 2014 to 2023.

Study Characteristics and Quality Assessment

Table 3 summarizes the characteristics of the studies who met the eligibility criteria. All seven (7) studies employed digital games as interventions for language learning, with three (3) of them detailing the development and creation of the games using various software tools (Game Maker Studio 8.1 and Genially) and frameworks (e.g., Octalysis). Among these, Wu (2018) and Roohani and Vincheh (2021) utilized the mechanics of a pre-existing game (Block Elimination Game and Phrasal Nerds: Phrasal Verbs, respectively). Wu (2018) also incorporated the ARCS model (Attention, Relevance, Confidence, and Satisfaction) in developing a mobile game-based English vocabulary practice system in the study's intervention group.

The studies who satisfied the inclusion criteria targeted different languages: five (5) focused on English, one (1) on Japanese, and one (1) on Chinese. All studies assessed the impact of game-based learning on vocabulary acquisition, while Cabrera-Solano (2022) and Roohani and Vincheh (2021) also examined its effectiveness in enhancing grammar skills. Additionally, all studies used self-made questionnaires as assessment tools for acquiring pre- and post-test scores that are validated by language professionals and underwent pilot testing. These questionnaires included various item types such as true or false, multiple choice, fill-in-the-blank, cloze tests, matching, rearranging, and sentence construction questions. Chen et al. (2023), Cabrera-Solano (2022), and Tang and Taguchi (2021) also conducted

Table 3 *Study Characteristics and Quality Appraisal of Studies Meeting the Review Eligibility Criteria (n = 7)*

Study	Non-Game-Based Learning used	Type of game used (software/ framework)	Duration (Weeks)	Population size (N)[a,b]	Age range	Language (skill)	Assessment tool	Score[c]
Zafar et al., 2014	Face-to-face classroom discussions	Digital (Game Maker Studio 8.1)	2	44[a]	19–21	English (Vocabulary)	30 Self-made T/F and MCQ questions	60%
Chen et al., 2023	Self-learning through video materials	Digital (Octalysis)	2	54[a]	25–47	Japanese (Vocabulary)	50 questions from a 900-question bank adopted from JLPT	80%
Cabrera-Solano, 2022	Online learning materials and virtual tutoring sessions	Digital (Genially)	16	61[b]	22–56	English (Grammar and Vocabulary)	Self-made questions	60%
Wu, 2018	Face-to-face classroom discussions	Digital (Block Elimination Game)	18	62[b]	University Students	English (Vocabulary)	50 items (100 points) self-made questions consisting of MCQ, fill-in-the-blank, and cloze test	80%
Tang & Taguchi, 2021	Self-learning through online materials	Digital (Questaurant)	2.3	48	17–27	Chinese (Pragmatics)	50 multiple-choice questionnaires consisting of 25 items of recognition test (scenario-based) and 25 dialogue completion tests (sentence formulation)	60%
Roohani & Vincheh, 2021	Face-to-face classroom discussions (Teacher-led)	Digital (Phrasal Nerds: Phrasal Verbs)	7	150	University EFL Students	English (Grammar and Vocabulary)	A 40-item questionnaire consisting of 15 items recognition of the correct particle, 10 items recognition of the main verb, and 15 items recognition of the whole phrasal verb	60%
Alhebshi & Gamlo, 2022	Face-to-face classroom discussions (Teacher-led)	Digital (Quizizz Application)	4	56	17–21	English (Vocabulary)	A self-made questionnaire that is based on the university's English course consisting of 25 multiple-choice questions that was administered through google forms	60%

Note: T/F = true or false, MCQ = multiple-choice questions, JLPT = Japanese Language Proficiency Test.
[a]Employed randomized controlled trial (RCT);
[b]Employed non-randomized controlled trial (NRCT);
[c]Studies were assessed using Hong, et al. (2018) Mixed Method Appraisal (NMAT) Tool in %.

additional interviews and questionnaires to examine the learning motivation of students who were exposed to game-based learning.

The duration of intervention among the studies varied greatly: two studies conducted the intervention over 16 to 18 weeks (Cabrera-Solano, 2022; Wu, 2018), while the remaining five completed it in a span of 2 to 7 weeks (Zafar et al., 2014; Chen et al., 2023; Tang & Taguchi, 2021; Roohani & Vincheh, 2021; Alhebshi & Gamlo, 2022). In terms of population size, most studies had participants ranging from 44 to 62, while Roohani and Vincheh (2021) had a larger sample size of 150 participants. This is due to the nature of the study, where they utilized 2 experimental and 1 control group (50 participants each) during their experimentation. Most studies reported participants whose ages fell within the pre-defined age range of 18 to 59 years (young to middle-aged adults). However, two studies (Wu, 2018; Roohani & Vincheh, 2021) did not explicitly report participant ages. For these studies, eligibility was determined through contextual inference based on the inclusion of university students in Taiwan and Iran, respectively, and supported by external sources describing typical age ranges of university students in those settings (Edara et al., 2021; Rahimisadegh et al., 2022). Edara et al. (2021) reported the age range of university students in northern Taiwan to be between 20 and 23, and Rahimisadegh et al. (2022) showed that Iranian university students are within the age range of 19 to 46 years old. Accordingly, these studies were included based on an inferential judgement rather than direct age reporting, and their compliance with age criterion should be interpreted cautiously.

Table 4 *MMAT Ratings of Each Study Meeting the Review Eligibility Criteria (n = 7)*

Studies	**Criteria from the Mixed Methods Appraisal Tool**									
	2.1	2.2	2.3	2.4	2.5	3.1	3.2	3.3	3.4	3.5
Zafar et al., 2014	0	1	1	0	1					
Chen et al., 2023	0	1	1	1	1					
Cabrera-Solano, 2022						0	1	1	0	1
Wu, 2018						0	1	1	1	1
Tang & Taguchi, 2021	1	1	0	0	1					
Roohani & Vincheh, 2021						0	1	1	0	1
Alhebshi & Gamlo, 2022						0	1	1	0	1

Studies were appraised using the Mixed Methods Appraisal Tool (MMAT) by Hong et al. (2018), and reporting was based on the authors' recommendations as well. A summary of the ratings for the studies who satisfied the inclusion criteria is presented in Table 4. Only categories 2 and 3 are shown in the reporting, as the studies were either quantitative RCT (category 2) or quantitative NRCT (category 3). No mixed method met the inclusion criteria for the full review, while quantitative-descriptive and qualitative studies were excluded from the meta-analysis. Most included RCTs lacked proper randomization (66.67%), as indicated by criteria 2.1, and failed to provide details on whether outcome assessors were blinded to the intervention (66.67%), as indicated by criteria 2.4. Additionally, all NRCTs lacked proper justification or methodology to prove that participants were representative of the target population (criteria 3.1), and most (75%) lacked appropriate control for confounders (criteria 3.4). Among the studies included, five (71.43%) were able to meet 3 criteria from the appraisal tool, while two (28.57%) were able to meet 80% of the criteria. Overall, the quality assessment of the studies who met the inclusion criteria yielded a grade between 60% and 80%, indicating a generally medium-to-high level of quality.

Data Extraction

Table 5 *Summary of Results of the Studies Meeting the Review Eligibility Criteria (n = 7)*

Study	Group	Sample size (n)	Pre-test score	p-value	Post-test score
Zafar et al., 2014	Non-Game-Based	22	16.22 ± 2.13	0.8291[a]	14.1 ± 2.25
	Game-Based	22	15.33 ± 2.33		15.95 ± 2.16
Chen et al., 2023	Non-Game-Based	27	85.04 ± 10.89	0.421	86.67 ± 9.46
	Game-Based	27	85.63 ± 9.01		90.44 ± 6.16
Cabrera-Solano, 2022	Non-Game-Based	29	75.57 ± 16.87	0.8959	78.35 ± 16.78
	Game-Based	32	88.46 ± 0.83		88.456 ± 0.83
Wu, 2018	Non-Game-Based	32	69.07 ± 4.34	0.276	69.17 ± 3.17
	Game-Based	30	68.38 ± 4.77		72.22 ± 4.45
Tang & Taguchi, 2021 (Recognition)	Non-Game-Based	24	14.17 ± 3.21	0.934	20.83 ± 3.74
	Game-Based	24	14.38 ± 3.92		22.12 ± 2.58
Tang & Taguchi, 2021 (Production)	Non-Game-Based	24	83.08 ± 15.48	0.773	92.69 ± 16.24
	Game-Based	24	85.52 ± 10.39		103.33 ± 13.89
Roohani & Vincheh, 2021	Non-Game-Based	50	13.74 ± 3.75	0.7483[a]	17.92 ± 3.38
	Game-Based	50	14.00 ± 4.32		23.68 ± 4.61
Alhebshi & Gamlo, 2022	Non-Game-Based	28	18.61 ± 5.38	0.1434[a]	15.43 ± 5.63
	Game-Based	28	16.39 ± 5.80		19.50 ± 4.44

Note: PV = Production Vocabulary, RV = Recognition Vocabulary; significant at $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.
[a]T-test was conducted to describe the difference between the baseline outcomes of the control and experimental groups.

The summary of pre- and post-test scores among the studies who satisfied the inclusion criteria is presented in Table 5. During the pre-test phase, the non-game-based and game-based learning groups generally exhibited overlapping score distributions, and most studies reported non-significant p-values from their corresponding t-tests. However, these baseline comparisons are presented for descriptive purposes only and should not be interpreted as evidence of group equivalence, particularly given the small sample sizes (22 to 50 participants) and the inclusion of non-randomized study designs. Baseline inferential statistics were included solely to identify potential imbalances between groups and not to establish comparability or to imply that confounding had been resolved. Three (3) of the included studies (Zafar et al., 2014; Roohani & Vincheh, 2021; Alhebshi & Gamlo, 2022) did not report baseline statistical analyses; therefore, t-tests were manually computed using the reported means, standard deviations, and sample sizes.

Risk of Bias

The risk of bias judgments for each study who satisfied the inclusion criteria is presented in Figure 2. According to RoB V2, among the three (3) randomized controlled trial studies (Figure 2a), the most common source of bias was deviation from the intended intervention and measurement of the outcome. This was primarily due to either the non-blinded assignment of participants or the absence of information regarding how the intervention was executed. However, all three studies demonstrated a low risk of bias due to missing outcomes, ensuring a complete study outcome and suggesting a reliable intervention effect estimate. For the non-randomized controlled trial studies, Figure 2b indicates that bias due to confounding was present in most studies, potentially stemming from the lack of statistical comparison between the baselines of the control and experimental groups. Nevertheless, all studies showed a low risk of bias due to missing data, reflecting a complete and proper execution of the intervention procedures.

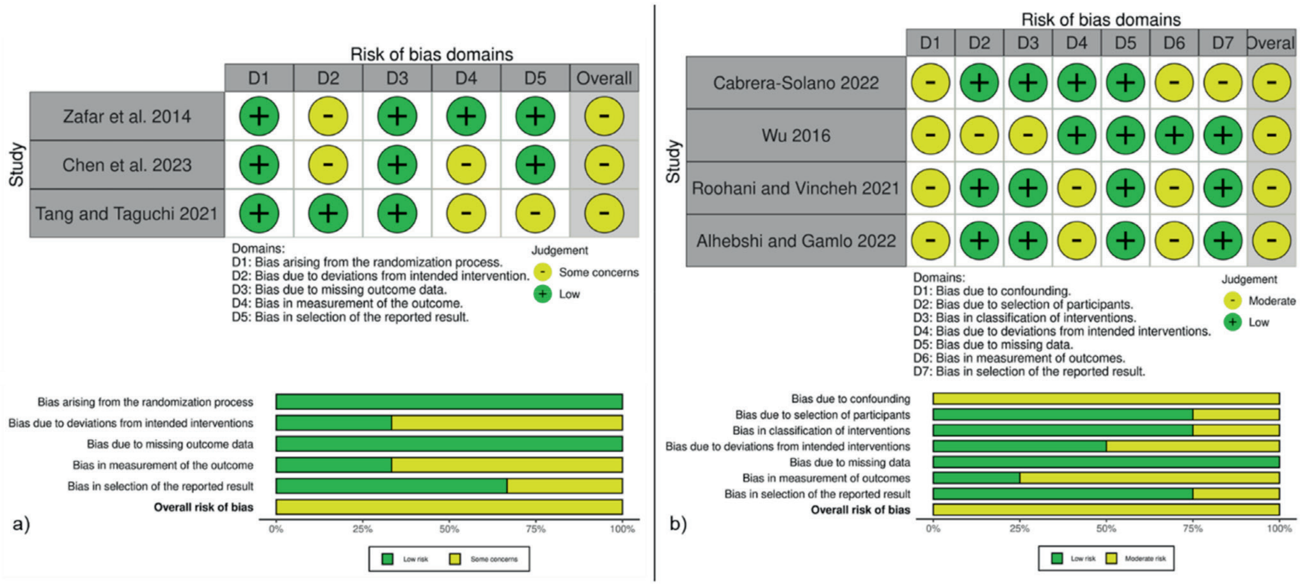

Figure 2 *Summary of Authors' Judgement on the Risk of Bias for Each Study Meeting the Review Eligibility Criteria. (a) Rob 2.0 for RCT and (b) ROBINS-I for NRCT*

Meta-Analysis

Table 6 presents the pooled SMD (Hedges' g) results of the sensitivity analysis using the one-study/subgroup-removal method. It should also be noted that, at this stage, the study of Cabrera-Solano (2022) was excluded due to unresolved inconsistency between the reported descriptive statistics and inferential results, limiting the included studies for the meta-analysis to $n = 6$. The pooled effect sizes of the six (6) included studies ranged from 0.81 to 1.05 and remained above the lower limit of the confidence interval across all single-study exclusions (low 95% CI = 0.59), while retaining a significant overall effect ($p < 0.01$). This suggests that no single study disproportionately influenced the overall findings of the meta-analysis. Subgroup sensitivity analysis was also conducted in order to assess whether weaker designs significantly influenced the overall effect size (Aung et al., 2026). As shown in Table 6, the pooled effect size remained above the lower limit of the 95% confidence interval (0.59) and maintained a statistically significant overall effect ($p < 0.01$). Subgroup analysis also remained above the lower confidence interval following individual exclusion of NRCTs (SMD = 0.68; 95% CI [0.19, 1.17], $Z = 2.70$, $p = 0.007$) or RCTs (SMD = 1.15; 95% CI [0.86, 1.44], $Z = 7.82$), with each removal retaining a statistically significant effect ($p < 0.01$). Finally, sensitivity analysis using alternative assumed correlations for the conservative computation of change scores yielded comparable pooled effects, supporting the stability of the combined outcome under reasonable correlation assumptions.

Table 6 *Sensitivity Analysis of the Studies Meeting the Review Eligibility criteria (n = 7)*

Sensitivity analysis	SMD (Hedges' g)	Lower limit	Upper limit	Z-value	p-value
Cabrera-Solano (2022)					
Inclusion (n = 7)	0.76	0.34	1.18	3.51	0.0004***
Exclusion (n = 6)	0.92	0.59	1.25	5.49	<0.00001***
One-study-removal					
Chen et al., 2023	1.05	0.77	1.33	7.36	<0.00001***
Zafar et al., 2014	0.87	0.50	1.25	4.54	<0.00001***
Wu, 2018	0.93	0.52	1.33	4.50	<0.00001***
Tang and Taguchi, 2021	0.99	0.62	1.35	5.34	<0.00001***
Roohani and Vincheh, 2021	0.81	0.49	1.14	4.90	<0.00001***
Alhebshi and Gamlo, 2022	0.87	0.49	1.26	4.42	<0.00001***
Subgroup analysis					
Exclusion of NRCT studies	0.68	0.19	1.17	2.70	0.007***
Exclusion of RCT studies	1.15	0.85	1.44	7.82	<0.00001***
Change score modification					
r = 0.3	0.79	0.51	1.06	5.58	<0.00001***
r = 0.5	0.92	0.59	1.25	5.52	<0.00001***
r = 0.7	1.17	0.74	1.660	5.37	<0.00001***

Note: SMD = standardized mean difference; significant at *p<0.1; **p<0.05; ***p<0.01. Of the seven eligible studies, six were retained in the primary meta-analysis, while one study (Cabrera-Solano, 2022) was considered only in sensitivity analysis.

The summary of the extracted and processed data, assumptions made during the meta-analysis, and the study-level effect sizes for each study meeting the inclusion criteria are shown in Table 7. Roohani and Vincheh's (2021) study involved three intervention groups (game-based, social media, and classroom-based), all of which are independent from each other. As specified in the methodology, only those intervention groups that are relevant to the study will be included in the meta-analysis hence, the social media group was excluded. This is due to the focus of the social media group, which is to enhance the vocabulary of the sample population through social interactions in mobile applications such as *Telegram* which is beyond the scope of the meta-analysis. Tang and Taguchi (2021), on the other hand, reported 2 different language outcomes from the same sample group. As indicated in the methodology used in this meta-analysis, multiple outcomes that were derived from the same sample group will be treated as statistically dependent and will be combined into a single effect size; therefore, the language learning outcomes of Tang and Taguchi (2021) were combined. A conservative correlation of r = 0.5 was used, and sensitivity analysis using r = 0.3 and r = 0.7 revealed a comparable result.

Figure 3 displays the meta-analysis results from the six included studies, comprising a total of 364 participants (183 in the non-game-based group and 181 in the game-based group). The study weights were generally distributed within a comparable range of 14.0% to 20.0% of the total weight. Comparable weight is supported by the results of Table 6 indicating that no single study overwhelmingly dominated the meta-analysis. Furthermore, the analysis revealed a large and significant effect size (SMD = 0.92, 95% CI [0.59, 1.25], Z = 5.49, p < 0.00001) favoring game-based learning over non-game-based learning methods in improving language learning outcomes. Based on the reported confidence interval, the

Table 7 *Summary of the Extracted and Processed Data, Assumptions, and Study-level Effect Size of the Studies Meeting the Review Eligibility Criteria*

Study	Outcome used	Group	Sample size (n)	Change score (Mean ± SD)	Combined outcomes (Yes/No)	Assumptions made	Included in the primary analysis? (Yes/No)	Study-level effect size [95% CI]
Chen et al., 2023	Grammar and Vocabulary	Non-Game-Based	27	1.63 ± 10.25	No	–	Yes	0.34 [-0.20, 0.88]
		Game-Based	27	4.81 ± 7.98				
Tang & Taguchi, 2021	Pragmatics	Non-Game-Based	24	8.13 ± 8.94	Yes	Multiple outcomes were treated as statistically dependent. Recognition and Production scores were combined into a single effect size	Yes	0.57 [-0.01, 1.14]
		Game-Based	24	12.78 ± 7.12				
Zafar et al., 2014	Vocabulary	Non-Game-Based	22	-2.12 ± 2.19	No	–	Yes	1.21 [0.56, 1.86]
		Game-Based	22	0.62 ± 2.25				
Alhebshi & Gamlo, 2022	Vocabulary	Non-Game-Based	28	-3.18 ± 5.51	No	–	Yes	1.15 [0.58, 1.72]
		Game-Based	28	3.11 ± 5.25				
Roohani & Vincheh, 2021	Grammar and Vocabulary	Non-Game-Based	50	4.18 ± 3.58	No	Exclusion of non-relevant intervention group and inferential assessment of age eligibility of the participants.	Yes	1.35 [0.91, 1.78]
		Game-Based	50	9.68 ± 4.47				
Wu, 2018	Vocabulary	Non-Game-Based	32	0.10 ± 3.89	No	Inferential assessment of age eligibility of the participants.	Yes	0.87 [0.34, 1.39]
		Game-Based	30	3.84 ± 4.62				
Cabrera-Solano, 2022	Grammar and Vocabulary	Non-Game-Based	29	1.79 ± 16.82	No	–	No	-
		Game-Based	32	0 ± 0.83				
OVERALL		**Non-Game-Based**	**183**					**0.92 [0.59, 1.25]**
		Game-Based	**181**					
							Prediction Interval	**[–0.04, 1.88]**

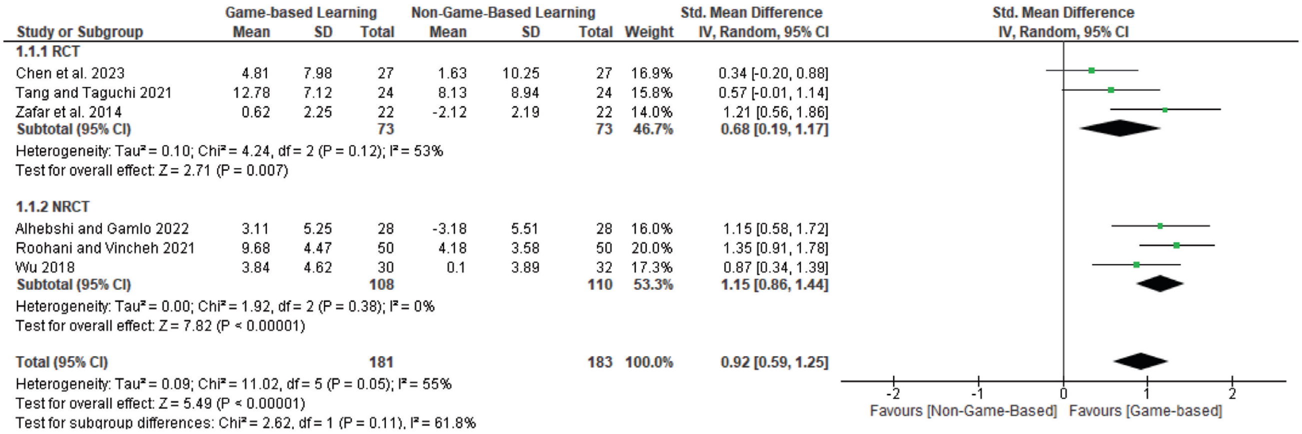

Figure 3 *Random-effects Model Forest Plot of the Difference Between Non-Game-Based (left) and Game-Based (right) Learning*

standard error of the pooled effect estimate (SE_m) was calculated to be 0.17. This value, together with the between-study variance ($\tau^2 = 0.09$), was subsequently used to calculate the prediction interval, yielding a range of –0.04 to 1.88.

The heterogeneity estimate ($I^2 = 55\%$) indicated moderate to substantial variability among the included studies. This was further examined using the Cochran's Q test ($Q = 11.02$, $p = 0.05$), which failed to reach conventional statistical significance. However, given the small number of studies included, this borderline result should be interpreted cautiously and should not be considered evidence of homogeneity. Accordingly, heterogeneity was interpreted in conjunction with the I^2 statistic and the prediction interval. To further examine the pooled estimate, fixed-effect and random-effect models were compared, as recommended by Page et al. (2024). Both models yielded similar and statistically significant effect sizes of 0.92 (random-effect) and 0.94 (fixed-effect), indicating that the overall estimate was not substantially altered under alternative pooling assumptions. Finally, subgroup differences were also assessed using the chi-square test ($chi^2 = 2.62$, $p = 0.11$), which did not indicate a statistically significant difference between subgroup categories. Nonetheless, this result should also be interpreted cautiously given the limited number of studies which limits the statistical power to detect true subgroup differences. Overall, these findings suggest that the pooled result was directionally consistent across model specifications and sensitivity analyses, although the certainty and generalizability of the estimate remain limited by the small and heterogeneous evidence base.

Discussion

The current review highlights the potential of Game-Based Learning (GBL) as a supplementary approach for addressing challenges in non-game-based language learning. This was examined through studies comparing GBL with conventional instructional methods, including printed classroom materials, teacher-led lectures, and classroom-based discussions. Several included studies reported improvements in language-related outcomes under GBL conditions. For example, Zafar et al. (2014) found that incorporating native language tips into a GBL strategy improved vocabulary and grammar retention among Saudi Arabian university students. Similarly, Chen et al. (2023) reported measurable gains in Japanese language skills using a game-based learning system developed through the Octalysis framework. Cabrera-Solano (2022) also reported significant improvements in grammar and vocabulary among EFL learners. Some of these studies additionally described positive learner experiences, such as increased motivation, engagement, or immersion; however, these outcomes were not quantitatively synthesized in the present meta-analysis and should therefore be interpreted as narrative observations only. Future research should examine the extent to which these broader learner-related factors contribute to language learning outcomes using more consistent and directly comparable measures.

Abdul Ghani and Wan Daud (2023) reported that computer game interventions were associated with improved Arabic communication skills among non-native speakers. Participants in the GBL group also expressed higher confidence, critical thinking, and motivation, and performed better compared to the group that utilized conventional learning methods. Wu (2018) also reported that incorporating the ARCS (Attention, Relevance, Confidence, Satisfaction) model into a mobile game-based vocabulary practice system was associated with improved English language proficiency among the participants. However, the author emphasized the need for features such as instant feedback, well-specified challenges, and multimedia integration in order to maintain participant interest and support knowledge retention (Wu, 2018). These studies collectively suggest that GBL may offer a potentially useful supplementary alternative to conventional learning methods, particularly for targeted language-related outcomes associated with language proficiency, although broader pedagogical implications should be interpreted cautiously given the limited and heterogeneous evidence base.

One important aspect of the inclusion and exclusion criteria for this review was the requirement for studies to report both pre- and post-test scores. This criterion allowed the computation of change scores for the meta-analysis. Change scores were used to account for observed pre-post differences within groups and to partially reduce the influence of baseline imbalance, particularly in NRCTs (Deeks et al., 2022). However, this approach does not establish baseline equivalence and does not eliminate the possibility of residual confounding inherent to non-randomized designs. Accounting for baseline differences is important when analyzing non-randomized studies, particularly because pre-existing group differences may influence post-intervention outcomes (Wuthrich et al., 2024). Although Cabrera-Solano (2022) reported baseline scores, it is notable that the pre- and post-test values for the game-based learning group were identical in both mean and standard deviation. Despite this, Cabrera-Solano (2022) reported no significant difference between the pre-test scores of the experimental and control groups (p-value of 0.8989, t-test = 0.1314). However, reanalysis of the published pre-test values indicated a statistically significant difference between groups (p-value = 0.0002), revealing an unresolved inconsistency between the reported descriptive statistics and inferential results. Due to this discrepancy, the study was excluded from the primary meta-analysis. Sensitivity analysis also showed that its exclusion yielded a higher pooled SMD (Hedges' g), although the overall direction and statistical significance of the pooled result remained unchanged (Table 6).

Table 4 indicates a specific concern within category 2.4 of the MMAT, reflecting a lack of information regarding blinding of outcome assessors among the included RCT studies. This is further corroborated by the RoB 2.0 assessment (Figure 2a), which shows "some concerns" regarding bias due to deviations from intended interventions (Domain 2). Additionally, potential bias may arise from the awareness of those delivering the interventions on the participants assigned groups, as reflected in the rating of "some concerns" in Domain 4 (bias in measurement of the outcome). The absence of explicit information confirming the blinded nature of the outcome assessments among the included RCT studies contributed to these concerns. These findings suggest that the lack of blinding among those delivering the intervention or conducting the assessment may have influenced the reported results (Sterne et al., 2019). As for the assessment of the included NRCTs using ROBINS-I, the lack of information regarding the blinding of outcome assessors also contributed to "some concerns" in the measurement of outcomes (Domain 6), as shown in Figure 2b. In addition, the "some concerns" rating for bias due to confounding (Domain 1) across the included NRCTs reflects the inherent risk of systematic baseline differences and uncontrolled confounding in non-randomized designs, rather than the absence of baseline significance testing itself. Although baseline characteristics may be described, statistical comparisons at baseline are descriptive only and should not be interpreted as evidence of group equivalence. Instead, the key concern is whether important baseline imbalances or other confounding variables were adequately identified, justified, and controlled for in the study design or analysis. In the included NRCTs, the lack of justification regarding whether samples were representative of the target population (category 3.1) and the absence of appropriate control for confounders (category 3.4) contributed to these concerns. These findings suggest that, while outcome measurement procedures may have been acceptable, the possibility of residual confounding limits the internal validity of the NRCT evidence and reduces its comparability to randomized trials (Sterne et al., 2016). Therefore, the integration of both study designs within the meta-analysis should be interpreted cautiously. This informed the decision to conduct a subgroup analysis to determine whether the inclusion of RCTs or NRCTs disproportionately influenced the overall effect size. As shown in Table 6, the removal of NRCT studies resulted in a lower overall effect size (SMD = 0.68), while the removal of RCTs led to a higher estimate (SMD = 1.15). Although a notable discrepancy between these estimates was observed, both values remained within the 95% confidence interval of the primary analysis, and Figure 3 revealed a non-significant subgroup difference ($chi^2 = 2.62$, $p = 0.11$). These findings suggest that while the inclusion of diverse study designs influenced the magnitude of the result, the overall direction and significance of the effect remained stable.

Previous studies have reported that GBL was associated with better language-related outcomes among young to middle-aged adults compared to non-game-based learning methods such as chalk and talk and conventional lecture classes (Zafar et al., 2014; Chen et al., 2023; Cabrera-Solano, 2022; Wu, 2018; Ghani & Daud, 2023). The earliest publication included in this review was from 2014, which is consistent with the reports of recent literature reviews (Schöbel et al., 2021; Noroozi et al., 2020). Video and computer game research began to rise around 2013 and reached its peak in 2019 (Schöbel et al., 2021). This trend is consistent with the findings of Noroozi et al. (2020), who documented the emergence of GBL in argumentation research way back in the 2000s with a rise in its publication from 2013 to 2018. Applications of GBL on other subject matter, such as mathematics and science, were also documented to rise between 2017 and 2019 (Yasir et al., 2021). Although the effect size was found to be significant and favors GBL, comparative studies regarding GBL and NGBL methods in other subject areas of knowledge may help determine whether similar patterns are observed across other domains. It should be noted that all included studies utilized digital GBL with no integration of non-digital games such as board games, card games, or role-playing activities. This could be due to the broader applicability, high controllability, and high adaptability of digital games, which can be further adjusted according to the available resources, unlike the specificity and elaborate nature of analog games (Pan et al., 2021). Although not addressed in this review, the utilization of analog or non-digital games in an educational setting has already been reported in previous studies (Asspun & Thummaphan, 2023; Nguyen et al., 2024). Sousa et al. (2023) highlighted the efficacy of analog games in improving students' cognitive and psychological outcomes, while Asspun and Thummaphan (2023) and Nguyen et al. (2024) showed how non-digital games significantly promoted engagement and motivation to learn among pediatric populations. However, it is important to note that most of these studies did not employ standardized assessment tools and primarily focused on populations beyond the scope of the present review. Additionally, all studies retained in the meta-analysis used digital games, and none that employed analog games met the inclusion criteria; hence, no analog games were reported. Future research may benefit from exploring a wider range of populations and incorporating both digital and analog games to better evaluate the overall effectiveness of game-based learning interventions.

English was the most frequently studied language among the studies included in this review. Five out of seven studies utilized English as the language of focus, which is similar to the results of Bylund et al. (2024), who reported English as the most frequent language used in studies concerned with second language acquisition. Several studies also referred to English as a second or foreign language (ESL or EFL), similar to the findings of previous studies (Zafar et al., 2014; Cabrera-Solano, 2022; Wu, 2018; Roohani & Vincheh, 2021; Alhebshi & Gamlo, 2022). Aside from variation in the language of focus, studies who met the review eligibility criteria also differ in several aspects of their interventions. Cabrera-Solano (2022) and Wu (2018) utilized longer intervention periods, while Zafar et al. (2014), Chen et al. (2023), and Tang and Taguchi (2021) employed shorter intervention periods of around two weeks. Cabrera-Solano (2022) and Roohani and Vincheh (2021) assessed both grammar and vocabulary, whereas other studies focused primarily on vocabulary learning and retention. Although the included studies explored GBL in second language acquisition, Chen et al. (2023) and Tang and Taguchi (2021) focused on languages other than English (Japanese and Chinese, respectively). These variations were one of the reason a random-effects model was used in the meta-analysis, as it allows for between-study variability arising from differences in intervention features, outcome measures, and study characteristics (Schünemann et al., 2023; Heung & Chiu, 2025).

Sensitivity analysis (Table 6) showed that the overall pooled effect remained statistically significant even after excluding specific subgroups or individual studies, suggesting that the overall direction of the finding was generally consistent across the tested scenarios. This may partly reflect the relatively comparable weights of the included studies (Figure 3), indicating that no single study appeared to dominate the pooled estimate (Clephas & Heesen, 2022). However, although statistical significance was retained after excluding either RCTs or NRCTs, the pooled Hedge's g shifted from large to moderate, suggesting that study design may have influenced the magnitude of the observed effect. This is consistent with Yao et al. (2025), who noted that including NRCTs alongside RCTs may increase heterogeneity and effect estimates. Accordingly, the combined inclusion of RCTs and NRCTs should be interpreted cautiously. Heterogeneity was also assessed using the I^2 statistic and Cochran's Q test. The I^2 value of 55% indicates moderate to substantial between-study variability (Clephas & Heesen, 2022). Cochrane's Q test was additionally examined to assess whether between-study variability exceeded what would be expected by chance alone (Chang et al., 2022).

In this study, a non-significant Cochran's Q was obtained ($p = 0.05$). However, in a meta-analysis with a small number of included studies ($n = 6$), Cochran's Q test is known to be underpowered, and a non-significant result does not definitively confirm the absence of heterogeneity (von Hippel, 2015). Furthermore, von Hippel (2015) found that in a meta-analysis including seven studies with no heterogeneity, I^2 tends to overestimate heterogeneity by about 12%, whereas in a meta-analysis of seven studies with over 80% heterogeneity, I^2 may underestimate it by approximately 28%. Due to the statistical limitations of I^2 and Cochran's Q, the authors cannot definitively claim that there is no meaningful heterogeneity. To better account for uncertainty and the potential range of effects in future settings, a 95% prediction interval was calculated, ranging from -0.04 to 1.88. Although the interval slightly includes the null value (zero), the range is predominantly positive, indicating that while the average pooled effect is statistically significant, the effect of GBL in a future comparable study may still be favorable overall, though some variability in the outcomes should be expected (Spineli & Pandis, 2020). This approach provides a more conservative and realistic interpretation of heterogeneity than the Q or I^2 statistics alone. Considered alongside the risk of bias assessment (Figure 2), these findings suggest that the pooled effect generally favored GBL across the included studies, although the limited number of studies and remaining between-study variability warrant cautious interpretation. As such, the results should not be taken as definitive evidence of stability or reliability, but rather as an indication that the observed favorable association between GBL and language-related outcomes was not solely dependent on a single heterogeneity metric and remained generally consistent within the constraints of the included studies.

Conclusion

This meta-analysis suggests that game-based learning (GBL) was associated with better language-related learning outcomes among young and middle-aged adults compared with non-game-based learning (NGBL) approaches in the included studies. Using standardized mean differences (SMD, calculated as Hedges' g) based on change scores, the pooled analysis showed a statistically significant large overall effect (SMD = 0.92, $p < 0.00001$) favoring GBL. However, these findings should be interpreted within the scope of the synthesized language-related outcomes and in light of the methodological limitations of the included studies, including the limited number of studies in the meta-analysis, moderate to substantial heterogeneity, mixed study designs, and some eligibility decisions based on contextual inference rather than directly reported data. Although some individual studies reported positive observations related to learner motivation and engagement, these outcomes were not quantitatively synthesized in the present review and should therefore be interpreted as narrative findings rather than as conclusions directly supported by the meta-analysis.

Most included studies were also conducted in researcher-controlled intervention settings, which may limit the generalizability of the findings to more naturalistic educational environments where implementation conditions and contextual factors may vary. Within these limitations, the findings suggest that GBL may be considered a potential useful supplementary instructional approach for supporting targeted language learning outcomes in educational settings, although its broader practical application should remain tentative pending stronger and more consistently reported evidence.

A key limitation of this review is the limited number of studies included in the quantitative synthesis. Future research should expand database coverage and include grey literature to capture a broader evidence base. Another limitation of this review is that some included studies did not explicitly report participant ages. In such cases, eligibility was inferred from contextual indicators such as university enrollment and country-specific educational norms. Although this approach was applied cautiously, it introduces some uncertainty regarding strict adherence to the predefined age-based inclusion criteria. This review was also limited to young and middle-aged adults, excluding minors and older adults; therefore, future meta-analyses should examine whether effects of GBL differ across age groups. In addition, future studies should investigate broader learner outcomes, such as motivation and engagement, using more consistent and directly comparable measures, particularly in more naturalistic classroom contexts.

References

Al-Azawi, R., Al-Faliti, F., & Al-Blushi, M. (2016). Educational gamification vs. game based learning: Comparative study. *International Journal of Innovation Management and Technology*, 131–136. https://doi.org/10.18178/ijimt.2016.7.4.659

Alhebshi, A. A., & Gamlo, N. (2022). The effects of mobile game-based learning on Saudi EFL foundation year students' vocabulary acquisition. *Arab World English Journal, 13*(1), 408–425. https://doi.org/10.24093/awej/vol13no1.27

Aung, N. M., Jurak, I., Mehmood, S., & Axon, E. (2026). Sensitivity analysis in meta-analysis: A tutorial. *Cochrane Evidence Synthesis and Methods, 4*(1), e70067. https://doi.org/10.1002/cesm.70067

Barz, N., Benick, M., Dörrenbächer-Ulrich, L., & Perels, F. (2023). The effect of digital game-based learning interventions on cognitive, metacognitive, and affective-motivational learning outcomes in school: a meta-analysis. *Review of Educational Research, 94*(2), 193–227. https://doi.org/10.3102/00346543231167795

Brydges, C. R. (2019). Effect size guidelines, sample size calculations, and statistical power in gerontology. *Innovation in Aging, 3*(4), igz036. https://doi.org/10.1093/geroni/igz036

Bylund, E., Khafif, Z., & Berghoff, R. (2023). Linguistic and geographic diversity in research on second language acquisition and multilingualism: an analysis of selected journals. *Applied Linguistics, 45*(2), 308–329. https://doi.org/10.1093/applin/amad022

Cabrera-Solano, P. (2022). Game-based learning in higher education: The pedagogical effect of genially games in English as a foreign language instruction. *International Journal of Educational Methodology, 8*(4), 719–729. https://doi.org/10.12973/ijem.8.4.719

Caldwell-Harris, C. L., & MacWhinney, B. (2023). Age effects in second language acquisition: Expanding the emergentist account. *Brain and Language, 241*, 105269. https://doi.org/10.1016/j.bandl.2023.105269

Chang, Y., Phillips, M. R., Guymer, R. H., Thabane, L., Bhandari, M., Chaudhary, V., Wykoff, C. C., Sivaprasad, S., Kaiser, P., Sarraf, D., Bakri, S., Garg, S. J., Singh, R. P., Holz, F. G., & Wong, T. Y. (2022). The 5 min meta-analysis: understanding how to read and interpret a forest plot. *Eye, 36*(4), 673–675. https://doi.org/10.1038/s41433-021-01867-6

Chen, C., Ming-Chaun, L., & Kuo, C. (2023). A game-based learning system based on octalysis gamification framework to promote employees' Japanese learning. *Computers & Education, 205*, 104899. https://doi.org/10.1016/j.compedu.2023.104899

Chung, E., Subramaniam, G., & Dass, L. C. (2020). Online learning readiness among university students in Malaysia amidst COVId-19. *Asian Journal of University Education, 16*(2), 45. https://doi.org/10.24191/ajue.v16i2.10294

Clephas, P. R. D., & Heesen, M. (2022). Interpretation of meta-analyses. *Interventional Pain Medicine, 1*, 100120. https://doi.org/10.1016/j.inpm.2022.100120

Cochrane Collaboration, 2020. Review Manager (RevMan) [Computer program]. Version 5.4.

Deeks JJ, Higgins JPT, Altman DG, McKenzie JE, Veroniki AA (editors). Chapter 10: Chapter 10: Analysing data and undertaking meta-analyses [last updated November 2024]. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.5. Cochrane, 2024. Available from cochrane.org/handbook.

Dixon, D. H., Dixon, T., & Jordan, E. (2022). Second language (L2) gains through digital game-based language learning (DGBLL): A meta-analysis. *Language Learning & Technology, 26*(1), 125. http://hdl.handle.net/10125/73464

Edara, I. R., Del Castillo, F., Ching, G. S., & Del Castillo, C. D. (2021). Religiosity, emotions, resilience, and wellness during the COVID-19 Pandemic: A study of Taiwanese university students. *International Journal of Environmental Research and Public Health, 18*(12), 6381. https://doi.org/10.3390/ijerph18126381

Erikson, E. H. (1994). *Identity and the life cycle.* WW Norton.

Ghani, M. T. A., & Daud, W. a. a. W. (2023). The impact of digital game-based learning towards Arabic language communication. *Jurnal Komunikasi Malaysian Journal of Communication, 39*(1), 407–424. https://doi.org/10.17576/jkmjc-2023-3901-23

Goh, E., Al-Tabbaa, O., & Khan, Z. (2023). Unravelling the complexity of the video game industry: An integrative framework and future research directions. *Telematics and Informatics Reports, 12*, 100100. https://doi.org/10.1016/j.teler.2023.100100

Higgins JP, Li T, Deeks JJ (editors). Chapter 6: Choosing effect measures and computing estimates of effect [last updated August 2023]. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editor(s). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.5. Cochrane, 2024. Available from *cochrane.org/handbook.*

Hong QN, Pluye P, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, Gagnon M-P, Griffiths F, Nicolau B, O'Cathain A, Rousseau M-C, Vedel I. Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada.

Infurna, F. J., Gerstorf, D., & Lachman, M. E. (2020). Midlife in the 2020s: Opportunities and challenges. *American Psychologist, 75*(4), 470–485. https://doi.org/10.1037/amp0000591

Jain, S., Sharma, S. K., & Jain, K. (2019). Meta-Analysis of fixed, random and mixed effects models. *International Journal of Mathematical Engineering and Management Sciences, 4*(1), 199–218. https://doi.org/10.33889/ijmems.2019.4.1-018

Jitpaisarnwattana, N., & Saville, N. (2025). Autonomous learning and students' perceptions of automated writing evaluation as a tool to improve writing skills. *Technology in Language Teaching & Learning, 7*(3), 102826. https://doi.org/10.29140/tltl.v7n3.102826

Liao, Y. K. (2010). Game-based Learning vs. Traditional Instruction: A Meta-Analysis of Thirty-Eight Studies from Taiwan. In D. Gibson & B. Dodge (Eds.), *Proceedings of SITE 2010–Society for Information Technology & Teacher Education International Conference* (pp. 1491-1498). San Diego, CA, USA: Association for the Advancement of Computing in Education (AACE). Retrieved April 30, 2025, from https://www.learntechlib.org/primary/p/33570/

McGuinness, LA, Higgins, JPT. Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. Res Syn Meth. 2020; 1–7. https://doi.org/10.1002/jrsm.1411

Nguyen, T. T. K., Ngo, T. P. A., Pham, A. T., Nguyen, D. D., Cheng, P. H., & Nguyen, V. B. (2024). Let's play! Transforming STEM education with board games. *Eurasia Journal of Mathematics Science and Technology Education, 20*(8), em2494. https://doi.org/10.29333/ejmste/14920

Noroozi, O., Dehghanzadeh, H., & Talaee, E. (2020). A systematic review on the impacts of game-based learning on argumentation skills. *Entertainment Computing, 35*, 100369. https://doi.org/10.1016/j.entcom.2020.100369

Ogawa, K. (2025). Development of task-based language teaching in virtual reality: A case study. *Technology in Language Teaching & Learning, 7*(3), 102540. https://doi.org/10.29140/tltl.v7n3.102540

Page MJ, Higgins JP, Sterne JAC. Chapter 13: Assessing risk of bias due to missing evidence in a meta-analysis [last updated August 2024]. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editor(s). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.5. Cochrane, 2024. Available from cochrane.org/handbook.

Pan, L., Tlili, A., Li, J., Jiang, F., Shi, G., Yu, H., & Yang, J. (2021). How to implement game-based learning in a smart classroom? A model based on a systematic literature review and Delphi method. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.749837

Rahimisadegh, R., Sharifi, N., Jahromi, V. K., Zahedi, R., Rostayee, Z., & Asadi, R. (2022). Self-medication practices and their characteristics among Iranian university students. *BMC Pharmacology and Toxicology, 23*(1). https://doi.org/10.1186/s40360-022-00602-5

Ru, Y., Norlizah, H. C., Burhanuddin, N. A. N., Liu, H., & Dong, J. (2025). The correlation between mindfulness and problematic smartphone use: A meta-analysis. *Addictive Behaviors, 164*, 108272. https://doi.org/10.1016/j.addbeh.2025.108272

Schöbel, S., Saqr, M., & Janson, A. (2021). Two decades of game concepts in digital learning environments – A bibliometric study and research agenda. *Computers & Education*, 173, 104296. https://doi.org/10.1016/j.compedu.2021.104296

Schünemann HJ, Vist GE, Higgins JPT, Santesso N, Deeks JJ, Glasziou P, Akl EA, Guyatt GH. Chapter 15: Interpreting results and drawing conclusions [last updated August 2023]. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane handbook for systematic reviews of interventions version 6.5*. Cochrane, 2024. Available from cochrane.org/handbook.

Sigmundsson, H., Haga, M., Elnes, M., Dybendal, B. H., & Hermundsdottir, F. (2022). Motivational factors are varying across age groups and gender. *International Journal of Environmental Research and Public Health, 19*(9), 5207. https://doi.org/10.3390/ijerph19095207

Sousa, C., Rye, S., Sousa, M., Torres, P. J., Perim, C., Mansuklal, S. A., & Ennami, F. (2023). Playing at the school table: Systematic literature review of board, tabletop, and other analog game-based learning approaches. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1160591

Spineli, L. M., & Pandis, N. (2020). Prediction interval in random-effects meta-analysis. *American Journal of Orthodontics and Dentofacial Orthopedics, 157*(4), 586–588. https://doi.org/10.1016/j.ajodo.2019.12.011

Sterne, J. a. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., . . . Higgins, J. P. T. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*, l4898. https://doi.org/10.1136/bmj.l4898

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., . . . Higgins, J. P. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, i4919. https://doi.org/10.1136/bmj.i4919

Topal, I. H. (2025). PlayPhrase.Me: A multimedia corpus for foreign language education. *Technology in Language Teaching & Learning, 7*(1), 102405. https://doi.org/10.29140/tltl.v7n1.102405

Tran, T. X., Vo, T. T. T., & Ho, C. (2023). From academic resilience to academic burnout among international university students during the post-COVID-19 new normal: An empirical study in Taiwan. *Behavioral Sciences, 13*(3), 206. https://doi.org/10.3390/bs13030206

Von Hippel, P. T. (2015). The heterogeneity statistic I2 can be biased in small meta-analyses. *BMC Medical Research Methodology, 15*(1). https://doi.org/10.1186/s12874-015-0024-z

Wu, T. (2018). Improving the effectiveness of English vocabulary review by integrating ARCS with mobile game-based learning. *Journal of Computer Assisted Learning, 34*(3), 315–323. https://doi.org/10.1111/jcal.12244

Wuthrich, V. M., Dickson, S. J., Pehlivan, M., Chen, J. T., Zagic, D., Ghai, I., Neelakandan, A., & Johnco, C. (2023). Efficacy of low intensity interventions for geriatric depression and anxiety – A systematic review and meta-analysis. J*ournal of Affective Disorders, 344*, 592–599. https://doi.org/10.1016/j.jad.2023.10.093

Yao, M., Mei, F., Ma, Y., Qin, X., Huan, J., Zou, K., Li, L., & Sun, X. (2025). Including non-randomized studies of interventions in meta-analyses of randomized controlled trials changed the estimates in more than a third of the studies: evidence from an empirical analysis. *Journal of Clinical Epidemiology, 183*, 111815. https://doi.org/10.1016/j.jclinepi.2025.111815

Yasir, G. M., Shaheen, A., & Hafeez, M. (2021). A systematic review on digital game-based versus traditional learning approaches. *Global Social Sciences Review, VI*(IV), 124–135. https://doi.org/10.31703/gssr.2021(vi–iv).12

Zafar, B., Mueen, A., Awedh, M., & Balubaid, M. (2014). Game-based learning with native language hint and their effects on student academic performance in a Saudi Arabia community college. *Journal of Computers in Education, 1*(4), 371–384. https://doi.org/10.1007/s40692-014-0017-