Revista de Psicodidáctica 30 (2025) 500158

Revista de Psicodidáctica

www.elsevier.es/psicod

Original

Effect of game-based learning on students' mathematics high order thinking skills: A meta-analysis

Bambang Sri Anggoro[a,*], Andi Harpeni Dewantara[b], Suherman Suherman[a,c], Rosida Rakhmawati Muhammad[a], and Sari Saraswati[d]

[a] Department of Mathematics Education, Universitas Islam Negeri Raden Intan Lampung, Bandar Lampung, Indonesia
[b] Educational Research and Evaluation, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[c] Doctoral School of Education, University of Szeged, Szeged, Hungary
[d] Department of Mathematics Education, Universitas Hasyim Asy'ari, Jombang, Indonesia

ARTICLE INFO

Article history:
Received 17 May 2024
Accepted 10 October 2024
Available online 22 October 2024

Keywords:
Game-Based Learning
Mathematics Higher Order Thinking Skills
Meta-Analysis
Thinking Skills

ABSTRACT

Educational research trends demonstrate a growing interest in understanding the impact of games on mathematics achievement. However, only a limited number of studies have synthesized previous quantitative studies to investigate how game-based learning (GBL) influences the enhancement of students' high-order thinking skills (HOTS) in mathematics. To address this gap, this meta-analysis study aims to investigate the effectiveness of GBL on students' mathematics HOTS. The study analyzed 40 effect sizes from 13 empirical studies published from 2010 to 2024, gathered from electronic databases such as Scopus, Eric, and Pro-Quest. R software was employed for analyzing effect sizes, detecting publication bias, and conducting subgroup analyses. Using the k-means algorithm, two studies identified as outliers were excluded. After removing these outliers, the findings revealed a positive effect of GBL on students' mathematics HOTS, with an overall effect falling within the small category ($g = .134$, $p < .001$). Additionally, moderator variables, including educational level, measured thinking skills, continent, control treatment, and intervention duration, significantly influenced the improvement of HOTS using GBL, while the sample size factor showed no significant impact. All findings, limitations, and implications are discussed in this article.

Efecto del aprendizaje basado en juegos en las habilidades de pensamiento de orden superior en matemáticas de los estudiantes: Un meta-análisis

Palabras clave:
Aprendizaje Basado en Juegos
Habilidades de Pensamiento de Alto Nivel en Matemáticas
Metaanálisis
Habilidades de Pensamiento

RESUMEN

Las tendencias de investigación educativa muestran un creciente interés en comprender el impacto de los juegos en el logro en matemáticas. Sin embargo, solo un número limitado de estudios ha sintetizado investigaciones cuantitativas previas para investigar cómo el aprendizaje basado en juegos (ABJ) influye en la mejora de las habilidades de pensamiento de alto nivel (HOTS) de los estudiantes en matemáticas. Para abordar esta brecha, se ha realizado un estudio de metaanálisis con el objetivo de investigar la efectividad del ABJ en las HOTS matemáticas de los estudiantes. El estudio ha analizado 40 tamaños de efecto de 13 estudios empíricos publicados desde 2010 hasta 2024, recopilados de bases de datos electrónicas como Scopus, Eric y Pro-Quest. Se ha empleado el software R para analizar los tamaños de efecto, detectar sesgo de publicación y realizar análisis de subgrupos. Utilizando el algoritmo k-means, se han excluido dos estudios identificados como valores atípicos. Después de eliminar estos valores atípicos, los hallazgos han revelado un efecto positivo del ABJ en las HOTS matemáticas de los estudiantes, con un efecto

DOI of original article: https://doi.org/10.1016/j.psicod.2024.500158
* Corresponding author.
E-mail address: bambangsrianggoro@radenintan.ac.id (B.S. Anggoro).

https://doi.org/10.1016/j.psicoe.2024.500158
2530-3805/

general situado dentro de la categoría pequeña ($g = .134$, $p < .001$). Además, se ha encontrado que variables moderadoras, incluido el nivel educativo, las habilidades de pensamiento medidas, el continente, el tratamiento de control y la duración de la intervención, han influido significativamente en la mejora de las HOTS mediante el ABJ; mientras que el factor de tamaño de la muestra no ha mostrado un impacto significativo. Se han discutido todos los hallazgos, limitaciones e implicaciones en este artículo.

Introduction

Mathematical concepts are applicable across a wide spectrum of everyday scenarios, ranging from uncomplicated tasks like determining discounts, profit and loss (Suherman et al., 2020; Wibawa et al., 2022), extending to more intricate applications in diverse fields such as engineering, economics, pharmaceutics, and accounting (López-Díaz & Peña, 2021; Mumcu, 2018). Additionally, mathematics has been regarded as a fundamental subject and the foundation of science and technology (Yeh et al., 2019). Regarding those essential roles, the instructional approach in classroom mathematics should not only emphasize procedural comprehension or theoretical knowledge, but also aim to develop high-order thinking skills (HOTS) (Lovianova et al., 2022), particularly analytical thinking, among mathematics learners. This proficiency is not only advantageous academically but also holds substantial relevance in effectively addressing and solving real-world challenges.

Developing HOTS, such as problem-solving, reasoning, critical thinking, and computational thinking, is crucial for students. However, mathematics educators often encounter significant challenges in this endeavor (Ganapathy et al., 2017). One primary difficulty faced by teachers is the quest for engaging and effective methods to stimulate students' high-order thinking. These challenges arise from the abstract, logical, and systematic characteristics of mathematics, coupled with the complexity of its symbols and formulas (Acharya, 2017). The challenges in improving students are multifaceted, encompassing diverse students' needs arising from their varied initial abilities (Smith et al., 2022), mathematical beliefs, and practical constraints associated with implementing specific teaching methods (Yerizon et al., 2022).

One potential strategy to address these challenges is by integrating games into the learning process. Game-based learning (GBL) has garnered significant interest from both researchers and practitioners (Qian & Clark, 2016). The integration of games in learning is intriguing to explore, as most students naturally enjoy playing games (White & McCoy, 2019). GBL involves incorporating games as educational tools, immersing students in a dynamic and interactive learning environment. GBL has emerged as an innovative learning approach that can enhance students' motivation, emotional involvement and enjoyment (Hartt et al., 2020). GBL also emerges as an innovative approach with the potential to assist educators in elevating students' HOTS.

Several previous studies have yielded valuable insights into the impacts of GBL, highlighting its significant enhancement of students' educational engagement across diverse educational level (Jabbar & Felicia, 2015; Shu & Liu, 2019). GBL has potential to make learning enjoyable and interactive, fostering students' interest and engagement, including cognitive involvement (Deng et al., 2020). In GBL, students are not only challenged to progress to higher levels but also derive satisfaction and joy upon successful completion of assignments or game missions (Freitas et al., 2017). In a more specific context, integrating game in the learning process holds the potential to enhance advanced cognitive thinking skills such as reasoning, critical thinking, and problem-solving (Angelelli et al., 2023; Bourke, 2019). Several prior studies support the idea that GBL has the potential to improve mathematics thinking skills. Asri & Jamaludin (2022) highlighted the effectiveness of Scratch games in developing thinking skills. Multiple studies demonstrated the enhancement problem solving skills through GBL using mind games (Demirel & Yilmaz, 2019), architecture-themed epistemic game (Ke, 2019). Indriani et al. (2019) found that problem-based learning assisted by monopoly game improved critical thinking ability, while Izzati et al. (2022) showed enhanced divergent thinking with snakes and ladders games.

Despite various studies have explored the use of GBL in mathematics education, the findings remain diverse and, at times, contradictory. Some studies advocate for the significant positive impact of GBL on students' HOTS (Al-Absi, 2017; Izzati et al., 2022; Lu et al., 2023; Ma et al., 2023; Taja-on, 2019), while others suggest no substantial difference compared to traditional teaching methods (Emihovich et al., 2022; Lee et al., 2014). This incongruity prompts a fundamental question: how can we establish consistent conclusions regarding the effects of GBL on students' mathematics HOTS? To comprehensively address this question, a meta-analysis, which systematically synthesizes results from multiple studies, needs to be conducted.

Several previous meta-analysis studies have examined GBL, highlighting its promise as a pedagogical approach that effectively enhances students' cognitive development (Turgut & Temur, 2017; Wang et al., 2022) and learning outcomes (Bai et al., 2020; Barz et al., 2024). Some meta-analyses have specifically focused on students' HOTS, such as the effects of GBL on computational thinking (Lu et al., 2023; Ma et al., 2023), and critical thinking (Mao et al., 2022). However, these studies often address general thinking skills rather than focusing specifically on mathematics. On the other hand, there are meta-analyses that examine the effects of GBL in a mathematical context, but their focus is on general mathematics achievement rather than HOTS. For instance, GBL has been reported as an effective educational method for improving students' mathematics achievement in specific educational levels, such as K-12 (Byun & Joung, 2018) and Pre K-12 grade levels (Tokac et al., 2019). Nonetheless, research specifically investigating the impact of GBL on students' mathematics HOTS across various educational levels remains limited.

To fill this gap, this current meta-analysis aims to assess the effectiveness of GBL -compared to traditional non-GBL- on students' mathematics HOTS. By integrating studies with diverse research designs and outcomes, the study aims to provide a nuanced understanding of the effectiveness of GBL in enhancing students' mathematical HOTS, such as problem-solving, critical thinking, matheamtics reasoning, computational thinking, and other relevant cognitive thinking skills (Suherman & Vidákovich, 2022). Through this systematic examination, the study seeks to contribute valuable insights to educators and researchers, shedding light on the potential of GBL to elevate students' mathematics HOTS.

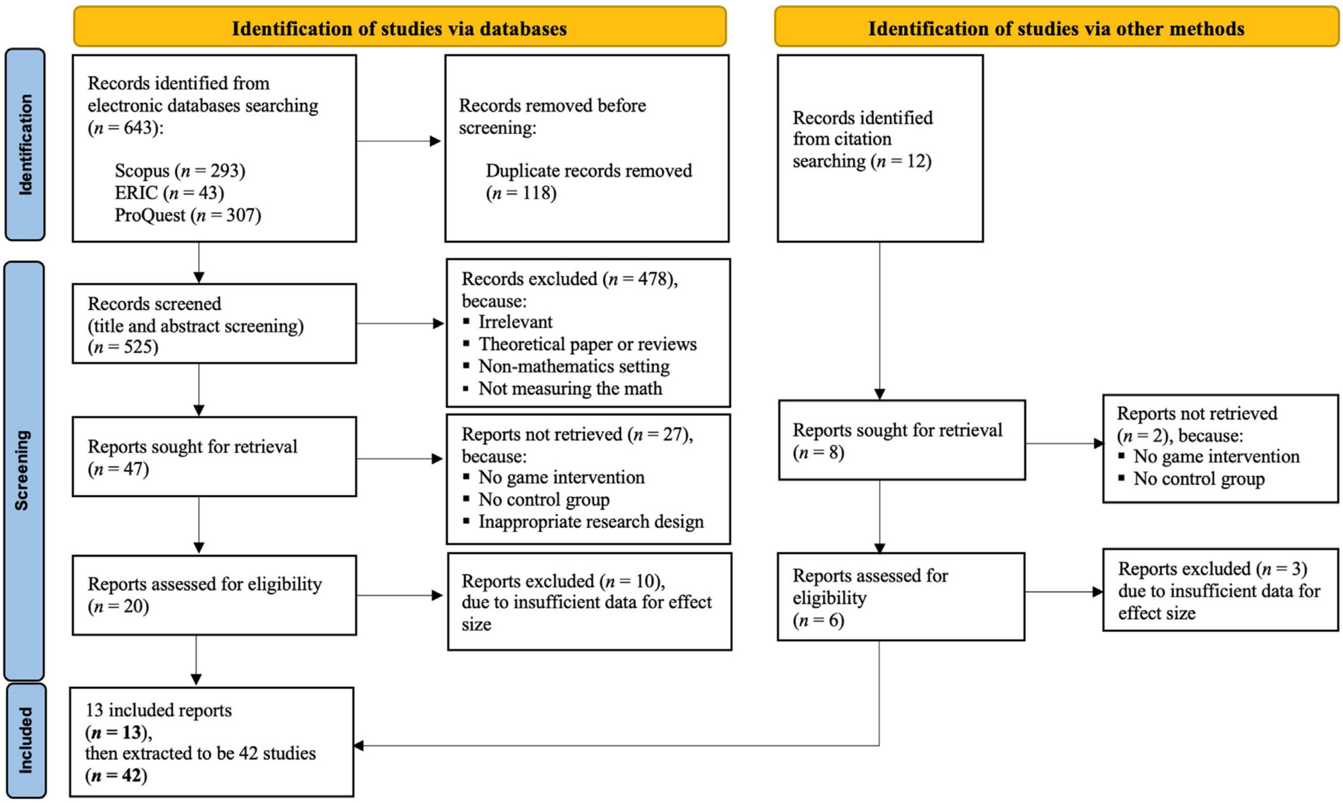

Figure 1. PRISMA flow diagram.

Method

This study employs a meta-analysis approach, a statistical method to synthesize findings from numerous quantitative studies within a particular research domain (Pigott & Polanin, 2020). In particular, this study applied a group contrast meta-analysis, involving the measurement of one or more variables across two or more respondent groups, followed by subsequent comparisons (Borenstein et al., 2009).

Literature search

To conduct a comprehensive meta-analysis about the effect of GBL on students' mathematics HOTS, this study employed the new Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) 2020 method (Page et al., 2021). The data sources consisted of studies (peer-reviewed journal articles and proceeding articles). Electronic databases, including Scopus, Eric, Pro-Quest, were queried for relevant studies published from January 2010 to January 2024. This period has seen significant advancements in educational technology and game design (Qian & Clark, 2016), increased adoption of game-based learning in educational settings (Plass et al., 2015), and a shift towards emphasizing higher-order thinking skills in mathematics curricula (Scherer & Beckmann, 2014). Additionally, research methodologies have become more sophisticated (Merchant et al., 2014), and educational policies have increasingly emphasized technology integration in mathematics education (Hsu et al., 2020). This timeframe thus provides a rich, recent body of research that reflects current educational contexts and practices, ensuring the meta-analysis findings are applicable to present-day mathematics instruction and future developments in game-based learning.

The search terms included variations of "game-based learning", "mathematics education," and "HOTS" in which the relevant keywords were combined with terms 'AND' and 'OR' as the Boolean operators (Wang et al., 2022). The first set of keywords encompassed terms *"game-based learning", "GBL", "educational game", "gamified learning", "gamification", "computer game", "digital game"*. The second set included keywords such as "*math**", "*math* learning*, "*math education*". The third set covered terms "*high-order thinking skill**", "*HOTS*", "*thinking skill*", "*cognitive thinking*", "*problem solving*", "*critical thinking*", "*creative thinking*", *"analytic thinking", "computational thinking", "reasoning*".

From the literature search, a total of 293 records were identified on Scopus, 43 on ERIC, and 307 on ProQuest (Figure 1). Furthermore, manual searching of literature was carried out, encompassing the examination of bibliographies of literature relevant to the meta-analysis topic. Following this, the acquired literature underwent screening based on the inclusion criteria set by the researcher.

Literature selection and inclusion criteria

Eight inclusion criteria have been established as guidelines for literature selection. Literature that does not meet all of the following inclusion criteria will be excluded. The eight inclusion criteria employed in this study are: (1) The literature must have been published between 2010 and 2024; (2) The literature should be published in peer-reviewed journals or conference proceedings, excluding theoretical papers, study reports, book reviews, and book chapters; (3) The studies should have been conducted in educational settings at all levels, from primary to higher education; (4) The studies needed to have full-text content in English. Studies not published in English were excluded; (5) The research should have a specific focus on the implementation of GBL interventions in mathematics classrooms; (6) The studies should report outcomes related to HOTS, covering problem-solving, critical thinking, computational thinking, creative thinking, mathematics reasoning, and other relevant cognitive abilities; (7) The studies should empirically investigate the impact of GBL on HOTS in mathematics learning (studies investigating the effects of GBL on students' HOTS in subjects other than mathematics were excluded); (8) The studies must have included at least one group using GBL and one control group with a traditional setting or non-game intervention (did not involve GBL). Studies without a control group were excluded; and, (9) The studies were required to provide sufficient data for calculating effect sizes. This includes reporting quantitative data such as means, standard deviations, and sample sizes for both the experimental and control groups or reporting t-tests, F-tests, or exact p-values along with sample sizes for both groups. Studies with insufficient data for calculating effect sizes were excluded.

Data extraction, coding scheme, and moderator variables

The data extraction and coding process in this meta-analysis adhered to a meticulous and systematic approach, encompassing various steps from literature search to final data analysis. The initial phase involved a literature search using specific keywords to identify relevant studies related to the impact of GBL on enhancing students' mathematics HOTS. Subsequently, literature selection was conducted based on predefined inclusion criteria.

Once the literature was successfully collected, a careful data extraction process was initiated from the selected studies. This process involved capturing information on study design, sample size, reported outcomes, and other essential variables. The meta-analysis analyzed thirteen empirical studies, including both journal articles and conference proceedings, generating 42 effect sizes for subsequent analysis. A summary of the literature search process is depicted in the new PRISMA flow diagram (Figure 1).

The selected thirteen literatures, which met the inclusion criteria, underwent a detailed data extraction process, resulting in the derivation of 42 effect sizes. Following this, a coding process was executed to identify the variables associated with the effect sizes, including the mean, standard deviation (*SD*), sample size (n), and effect sizes related to posttest scores for both experimental and control groups. Additionally, any moderator variables present in each study were identified during the coding process.

The codebook (coding scheme) serves as a comprehensive guide for coding each study, containing essential information such as: (1) The literature's details, including auhor(s)' last name, publication year, the title of the studies, and the type of publication (peer-reviewed articles or proceedings); (2) Sample size: small ($n < 30$) (S) and large ($n \geq 30$) (L). This sample size refers to the sample of experiment class in the analyzed studies; (3) Cognitive thinking skills measured: problem solving (PS), computational thinking (CoT), critical thinking (CrT), reasoning (Rs), and other (O); (4) Educational level : elementary school (ES), secondary school (SC), and undergraduate (UG). These educational level categories used in this meta-analysis were determined based on the classifications provided in the original articles, rather than specific age ranges. However, the age ranges for each study analyzed are presented in Table 1 (Appendix); (5) Control treatment: traditional (Tr) and multimedia (MM); (6) Continent : Asia (A), Europe (E), North America (NoA), and Other (O); (7) Intervention duration: < 10 hours (ID 1), 10 – 20 hours (ID 2), > 20 hours (ID 3), and not specified (NS); and, (8) Statistical information to calculate effect sizes, including total sample of experiment group (*NE*), average score of experiment group (ME), and standard deviation of experiment group (SD E), total sample of control group (NC), average score of control group (MC), and standard deviation of control group (SD C). Furthermore, two raters were involved in assessing the literature coding conducted by the researchers in this meta-analysis. To ensure reliability, inter-rater consistency was measured using the kappa statistics (Ho et al., 2019).

Effect size calculation and statistical models

This research compared mathematics cognitive achievement between control and experimental groups, then the effect size was calculated using the standardized mean difference (*SMD*) proposed by Hedges (1981). Descriptive statistics from most studies, including mean ($\overline{X}_1$, $\overline{X}_2$), standard deviation (s_1, s_2), and sample size (n_1, n_2) for both groups, allow us to compute *SMD* values using the following formula (Lipsey & Wilson, 2001):

$$d = \frac{\overline{X}_2 - \overline{X}_1}{S_{pooled}}$$

Where,

$$S_{pooled} = \frac{\sqrt{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2}}{n_1 + n_2 - 2}$$

Hedges (1981) found that the d value slightly overestimates the absolute population parameter, leading to a proposed modification known as *Hedges' g*. To convert from d to g, a correction factor called J is needed, calculated with the following formula:

$$J = 1 - \frac{3}{4df - 1}$$

where df is degrees of freedom calculated as:

$$df = n_1 + n_2 - 2$$

with the corrected effect size, Hedges' g was computed as:

$$g = J \times d$$

After calculating the effect size for each study, the aggregate effect (mean effect) was computed using a random effect (RE) model. The selection of a random effects model over a fixed model was based on the researchers' assumption that effect sizes in various studies could potentially stem from distinct populations, and these different populations have their own sampling distributions (Borenstein et al., 2009). This is reinforced by the opinion of Cooper (2016) that the diversity in experimental settings in each study (e.g., gender, country, education level) would be better suited for analysis using a random effects model.

In our analysis, we employed a frequentist framework, which relies on classical statistical methods to estimate parameters and infer population characteristics based on observed data. This framework was used to assess the pooled effect size and the influence of moderator variables. Furthermore, we used inverse variance weighting for the meta-analytic integration, as the most common way to calculate a pooled effect size (Harrer et al., 2021). This method gives more weight to studies with more precise effect size estimates, which are typically those with larger sample sizes or smaller variances. The overall effect size was calculated as a weighted average of the effect sizes from individual studies, where the weights correspond to the precision of each study's effect size estimate. This approach allows us to incorporate the variability both within and between studies, leading to a more accurate estimate of the overall effect size.

Heterogeneity among the effect sizes was assessed using the Q statistic and quantified with the I^2 index. The between-studies variance was estimated using the DerSimonian and Laird method. The corresponding confidence intervals for the heterogeneity estimates have also been included. This approach allows us to account for variability both within and between studies, providing a more accurate estimate of the overall effect size.

To calculate the effect size for each study and the aggregate effect, R software version 2023.9.0.463 (Posit Team, 2023) was employed with the 'meta' package (Balduzzi et al., 2019). The random effects model implemented in this package allows for the incorporation of variability both within and between studies, and it provides a robust estimate of the overall effect size. The frequentist approach used in this analysis involves calculating point estimates and confidence intervals to make inferences about the population effects.

Outlier detection

In meta-analysis, random effects are typically used to account for unexplained variation beyond within-study variability, but this method may not suffice when dealing with outlier studies (Beath, 2014). It is crucial to consider the identification of outliers (Riahi & Mokhayeri, 2017) because these outliers possess an effect size so extreme that it markedly diverges from the overall effect. The R software version 2023.9.0.463 (Posit Team, 2023) was utilized for detecting the outlier(s), employing the 'metafor' and 'dmetar' packages. The 'dmetar' package incorporates the 'find.outliers' function, implementing the outlier removal algorithm (Harrer et al., 2021). Additionally, he 'gosh.diagnostic' function, utilizing different cluster algorithms to identify data patterns, was also utilized (Harrer et al., 2021). In this study, the k-means algorithm (Hartigan & Wong, 1979) was specifically selected from these algorithms.

Moderator variable analysis

Analyzing moderator variables is crucial in meta-analysis, helping researchers understand the factors influencing variations in effect sizes among studies and informing the evaluation of current interventions and the potential formulation of more impactful ones (Li et al., 2020). In this study, the moderator variables to be analyzed include sample size, educational level, continent where the study was conducted, as well as control treatment and intervention duration. The exploration of research outcome diversity among various studies involved the utilization of a heterogeneity test (Q test). A noteworthy finding in the Q statistic indicates the potential consideration of each study as originating from a common population. Essentially, the substantial variations in the collective impact of each element within the moderator variable, highlighted by a significant Q statistic, underscore the potential significance of moderator variable analysis. Analysis of all moderator variables in this study was conducted using ANOVA-like models. Within these models, presentation of mean effects (g) within groups, 95% confidence intervals (CI), and heterogeneity between groups (Qb) was made. A noteworthy Qb statistic signals significant distinctions in aggregate effects among components of the moderator variable. The analysis of moderator variables was facilitated by R software version 2023.9.0.463 (Posit Team, 2023) with 'meta' package (Balduzzi et al., 2019).

Publication bias assessment

Publication bias evaluation aims to identify potential bias in the literature and its impact on the overall conclusions of the meta-analysis. To assess the publication bias in this meta analysis study, a three-pronged approach was adopted. The examination involved the implementation of funnel plots and Egger's regression test. The visual inspection of the funnel plot was employed as an initial step to discern any asymmetry. A symmetric distribution within the funnel plot was considered indicative of an absence of publication bias (Card, 2011). This graphical representation facilitated the identification of potential outliers and provided an overall visual impression of the publication bias landscape. Subsequently, Egger's regression test, recognized for its quantitative assessment of funnel plot asymmetry, was applied to rigorously evaluate the presence of publication bias. This statistical approach quantified the degree of asymmetry within the funnel plot, contributing a quantitative dimension to the assessment of bias. The evaluation aimed to discern whether there was a systematic relationship between the effect sizes and their precision, potentially indicating publication bias.

In addition to the above methods, we also applied the Trim and Fill method to further assess publication bias. This is one of the most common methods to adjust for funnel plot asymmetry (Duval & Tweedie, 2000). Trim and Fill method involves identifying potential missing studies from the funnel plot and adjusting the meta-analysis to account for these missing studies, thus providing a more accurate estimate of the effect size. This comprehensive approach allowed us to address potential biases more robustly and enhance the credibility of our findings.

To execute the analysis of publication bias, R software version 2023.9.0.463 (Posit Team, 2023) was utilized in conjunction with the 'meta' package (Balduzzi et al., 2019) and the 'metafor' package (Viechtbauer, 2010). The integration of these statistical tools enabled a comprehensive exploration of publication bias, reinforcing the robustness and credibility of the meta-analysis findings in elucidating the impact of game-based learning on students' mathematics analytic thinking skills.

Results

Study overview

In the course of the literature review and selection process, a total of thirteen relevant sources, comprising experimental studies investigating the impact of GBL on students' mathematical analytic skills, were identified. The synthesis of these 13 articles yielded a dataset encompassing 42 individual studies, each contributing a distinct effect size to the meta-analysis. The studies included in this meta-analysis meet the predetermined inclusion criteria: they are peer-reviewed journal articles or conference proceedings published in English between 2010 and 2024. These studies empirically investigate the impact of GBL on HOTS in mathematics education across all levels (from primary to undergraduate). The research design must be experimental, involving at least one group using GBL and one control group with traditional or non-game interventions (non-GBL). Additionally, the studies must provide sufficient data to calculate effect sizes.

The characteristics of the studies included in this meta-analysis, along with the detailed distribution of these studies, are presented in Appendix (Table 1). This table provides an overview of the diverse literature sources, publication types, measured thinking skills, intervention durations, sample sizes, educational levels, continents where the studies were conducted, and control treatments. A total of 42 studies are involved, comprising 40 journal articles and two conference proceedings, conducted across four different continents: Asia ($n = 12$), North America ($n = 3$), South America ($n = 1$), and Europe ($n = 22$). These studies examine the impact of GBL on HOTS in mathematics students at the following educational levels: primary school ($n = 29$), secondary school ($n = 6$), and undergraduate ($n = 5$). The research focuses on various aspects of HOTS, including problem-solving ($n = 3$), mathematical reasoning ($n = 18$), and other skills such as computational thinking, critical thinking, and divergent thinking ($n = 6$). In the experimental groups, sample sizes are categorized into two groups: small samples ($n < 30$) in 11 studies and large samples ($n \leq 3$) in 29 studies. The duration of the GBL intervention varies and is categorized into four groups: 12 studies with intervention durations of less than ten hours, three studies with durations both less than ten hours and more than 20 hours, and the remaining studies lack detailed duration information. In the control groups, teaching methods include

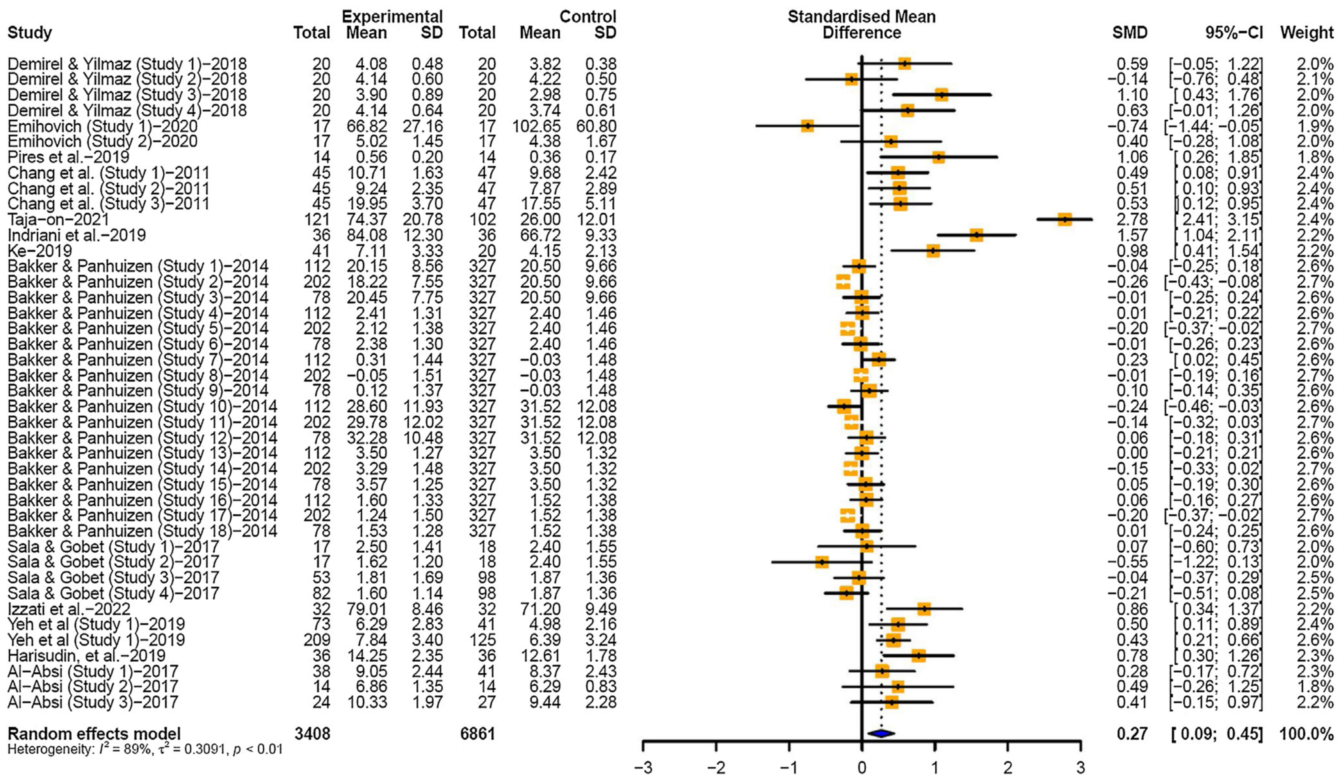

Figure 2. Forest Plot for 42 Studies (Pre-Outlier Detection).

multimedia-based instruction in 20 studies, with the rest utilizing traditional teaching methods.

To provide a comprehensive perspective on the included studies, an examination of key characteristics was conducted. These characteristics were organized into four primary categories: literature, study design, sample, and intervention (gaming) features. Within each category, subcharacteristics were identified, serving as moderator variables that provided a detailed perspective on the studies. In this study, two raters assessed the literature coding conducted by the researchers. Inter-rater consistency was measured using the kappa statistic, which yielded a kappa value of .78, indicating a 'substantial agreement' among the raters. It confirmed that the two raters' decisions were highly consistent, thereby supporting the reliability of the literature coding performed by the researchers (Ho et al., 2019).

Main analysis results before detecting outliers

The analysis utilizing the random effects model revealed that the average impact size across the 42 studies was .266 (p < .001), and the 95% confidence interval ranged from .086 to .446 (Figure 2). These findings indicate a statistically significant impact of GBL on students' mathematics HOTS compared to traditional (non-GBL) learning methods, although the aggregate effect size falls into the 'moderate' category, as defined by Cohen (2013), where effect sizes between 0.20 and 0.50 are considered moderate for experimental studies. Thus, the positive impact of implementing GBL on mathematics HOTS, compared to traditional methods, falls within the medium/moderate effect size category. Moreover, the analysis also revealed diverse effect sizes among the 42 studies (Q = 385.71, df = 41, p < .001), with a level of heterogeneity (I^2) reaching 89.4%. These findings indicate a considerable variation in the effect sizes employed in this study.

Outlier detection

Identifying outliers, data points that exhibit considerable deviation from the standard, can wield a noteworthy influence on the overall outcomes. Consequently, it is recommended to eliminate outliers and subsequently reassess the cumulative impact (Riahi & Mokhayeri, 2017). The plots from the K-means analysis (Figure 3) illustrate studies with data points positioned significantly distant from the horizontal line (studies 11 and 12), acknowledged as outliers. The results of the analysis uncover the existence of two studies identified as outliers, encompassing the research conducted by Taja-on (2019) and Indriani et al. (2019).

Summary effect analysis after removing outliers

After the exclusion of two studies identified as outliers, a reevaluation of the collective impact of the remaining 40 studies was undertaken. In these studies, the effect sizes (ES) range from -.74 to 1.10 (Figure 4). Among them, 15 studies (37.50%) depict a negative effect size, indicating that GBL was not more effective compared to traditional learning (non-GBL) in the control group. Conversely, 25 studies (62.50%) show a positive effect size, suggesting that students in the experimental group, implementing mathematics GBL, achieved higher mathematics HOTS than those in the control group with non-game approaches. To summarize, the results reveal an overall positive and significant effect size with g = .134 (p < .001, 95% CI = [.028; .239]) considerably lower than the original effect size before removing the outlier studies (g = .27), highlighting the importance of outlier detection in meta-analyses (Viechtbauer & Cheung, 2010). In contrast to the condition before excluding the outliers, the overall effect size now falls within the 'small effect size' category (Cohen, 2013). Figure 4 provides the forest plot, illustrating the studies' effect sizes and confidence intervals. Additionally, the heterogeneity test yielded a significant result (Q = 143.62, df = 39, p < .001), indicating variation in the effect sizes

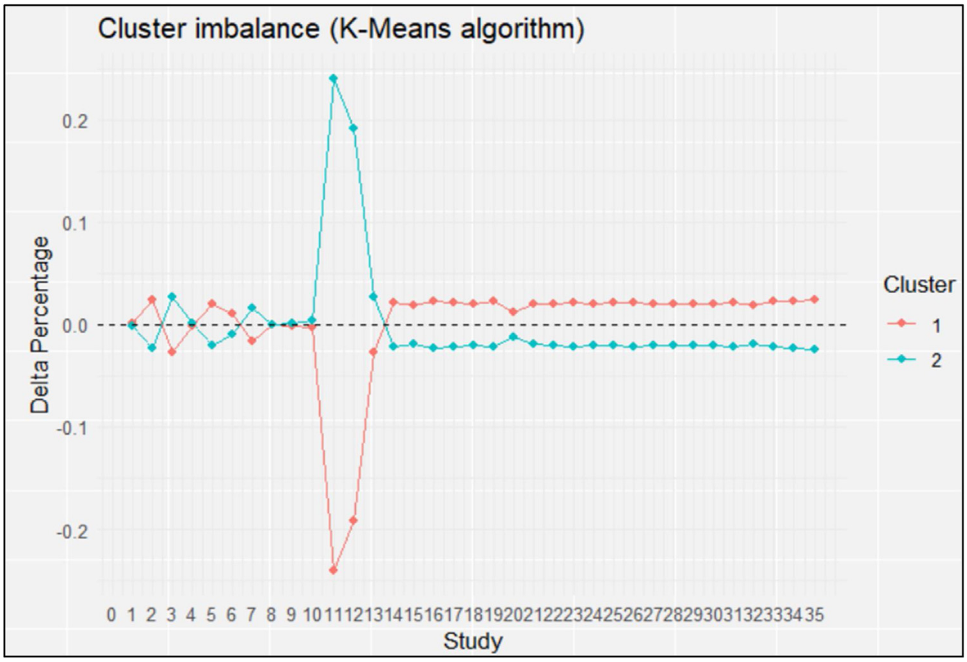

Figure 3. K-Means Plot for Detecting Outliers.

(I^2 = 72.8%). According to Higgins et al. (2003), an I^2 value above 75% suggests considerable heterogeneity, indicating that a substantial proportion of the observed variance reflects real differences in effect sizes rather than sampling error. This heterogeneity underscores the need for exploring potential moderators that might explain the variability in GBL effectiveness (Borenstein et al., 2009).

Publication bias evaluation

The funnel plot in Figure 5 visually represents the results of a publication bias assessment using the Trim and Fill method (Duval & Tweedie, 2000). This analysis revealed substantial publication bias in the meta-analysis. The asymmetry observed in the plot, with fewer studies on the left side, initially suggested missing studies with negative or null findings. The Trim and Fill method confirmed this, identifying 11 potentially missing studies. Crucially, after adjusting for these missing studies, the results changed dramatically. The originally reported effect size was significantly reduced and became statistically non-significant (*SMD* = -.019, 95% *CI* [-.154; .116], *p* = .781). This substantial shift suggests that publication bias may have inflated the original findings, a common issue in educational research (Pigott et al., 2013). This aligns with known challenges in educational research, where studies with positive or significant findings are more likely to be published, potentially skewing meta-analytic results. The funnel plot and associated analysis serve as a critical reminder of the importance of addressing publication bias in research synthesis and the need for caution when interpreting unadjusted meta-analytic findings in fields prone to such bias.

Interestingly, Egger's test (Egger et al., 1997) provided a contrasting result (t = 4.55, p = .082), not indicating significant publication bias. The associated funnel plot (Figure 6) shows a relatively symmetrical distribution of effect sizes. This discrepancy between the Trim and Fill results and Egger's test is not uncommon in meta-analyses and highlights the importance of using multiple methods to assess publication bias (Sutton et al., 2000). It also suggests that different types of publication bias might be at play, such as outcome reporting bias or time-lag bias (Dwan et al., 2008; Ioannidis & Trikalinos, 2005). The funnel plot (see Figure 6) reveals asymmetry, particularly a lack of studies in the lower left quadrant, suggesting potential publication bias in this meta-analysis. This asymmetry indicates that smaller studies with negative or null results may be underrepresented, a phenomenon known as the "file drawer problem" (Rothstein, 2005). The plot shows a trend where smaller studies (with larger standard errors) tend to report larger positive effects, while there's a noticeable absence of smaller studies reporting negative effects. This pattern aligns with Sterne et al. (2011) description of publication bias in funnel plots and supports the earlier findings from the Trim and Fill method. While the Egger test may not have indicated significant bias, visual inspection can reveal patterns that statistical tests might miss (Egger et al., 1997). This asymmetry underscores the need for cautious interpretation of the meta-analytic results and highlights the potential impact of publication bias in educational research, where small studies are common (Ioannidis & Trikalinos, 2007).

The substantial change in effect size after applying the Trim and Fill method underscores the critical importance of addressing publication bias in meta-analyses. As Ioannidis (2005) argues, publication bias can lead to an overestimation of effect sizes and potentially false-positive findings in the literature. In this case, the initial small positive effect of GBL on mathematics HOTS appears to be largely attributable to publication bias, suggesting that the true effect may be negligible or non-existent when considering the full landscape of conducted studies, including those potentially unpublished.

Moderator variable analysis

The analysis results of moderator variables, encompassing the measured thinking skills, educational level, sample size (Li et al., 2020), continent where study conducted, control treatment, and intervention duration (Prieto-Rodriguez et al., 2020), are depicted in Figure 7.

Thinking skills measured

The initial moderator variable 'thinking skills measured' consisted of four groups: *problem solving, critical thinking, mathematical*

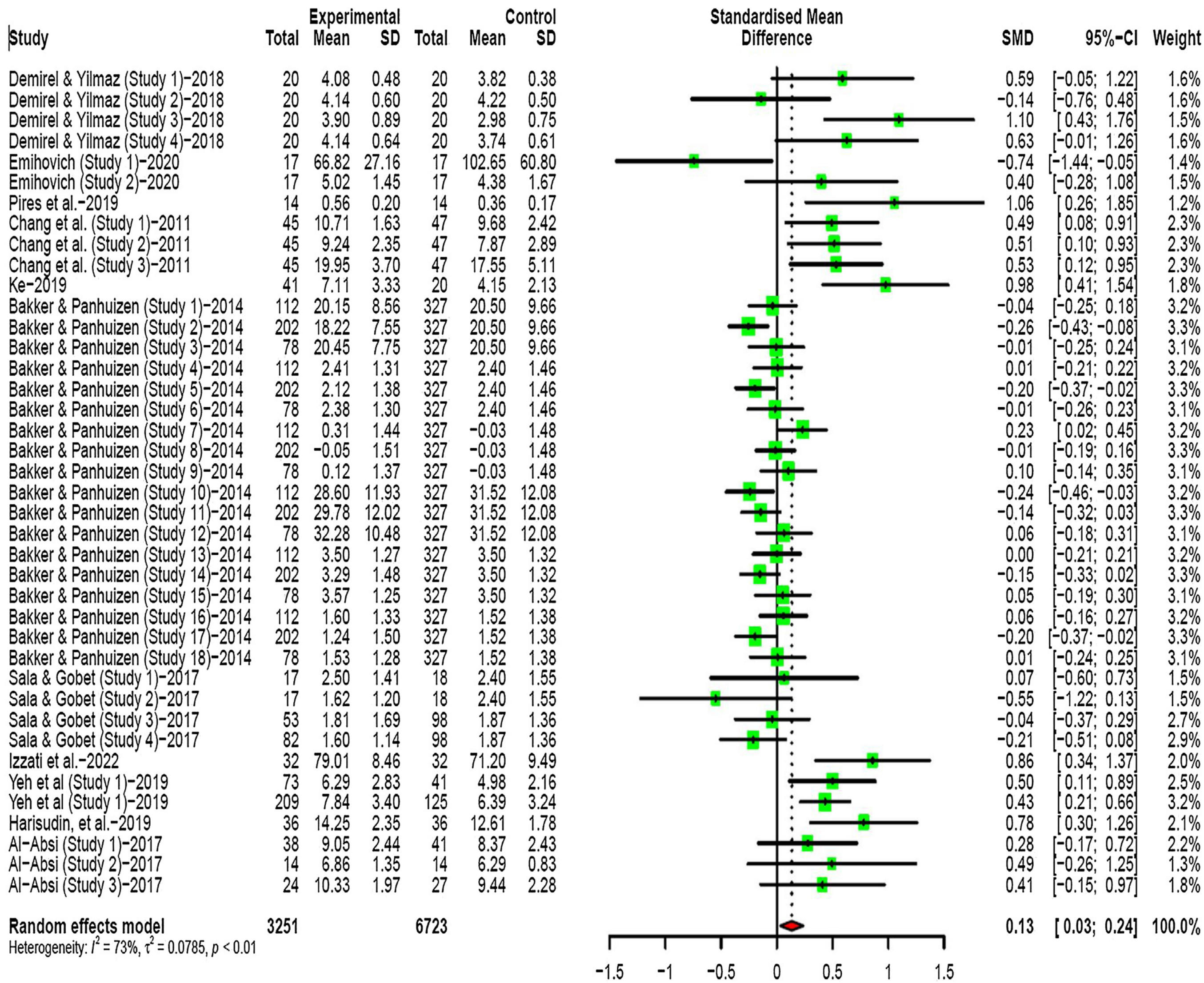

Figure 4. Forest Plot for 40 Studies (Post-Outlier Detection).

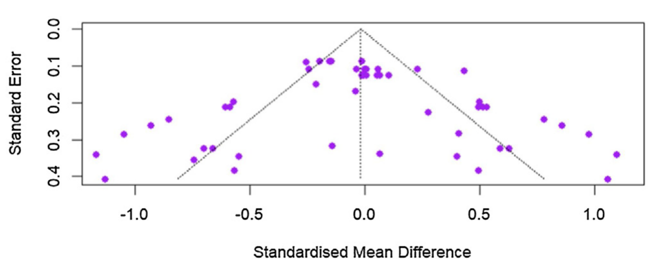

Figure 5. Funnel plot using fill and trim method.

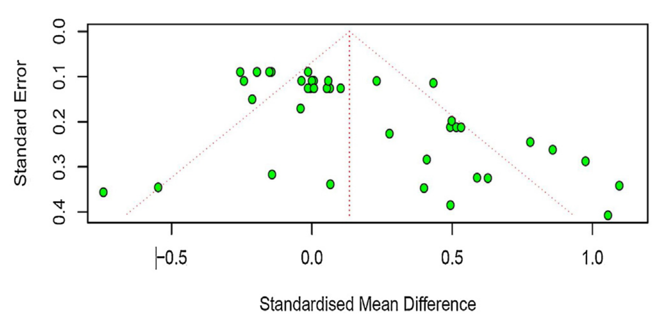

Figure 6. Funnel plot using Egger method.

reasoning, and other skills. However, after the removal of outliers, only three groups of thinking skills remained, as both studies removed belonged to the category of critical thinking. Overell, the analysis result show that there is a significant difference in the effectiveness of GBL based on the type of thinking skill measured (Qb = 32.63, p < .001). The analysis reveals that GBL has a positive yet small effect on enhancing students' *problem-solving skills,* with an effect size of g = .29. Among the three measured thinking skill groups, GBL demonstrated notable positive effectiveness for 'other skills,' such as *critical thinking*, *computational think*ing, and *divergent thinking*, with a larger effect size of g = .61. In contrast, for *mathematical reasoning skills* (with g = -.05 and 95% confidence interval of [-.12; .01] is not statistically significant, as the confidence interval includes zero. This indicates that GBL does not show a significant difference in improving students' *mathematical reasoning skills* compared to non-GBL interventions.

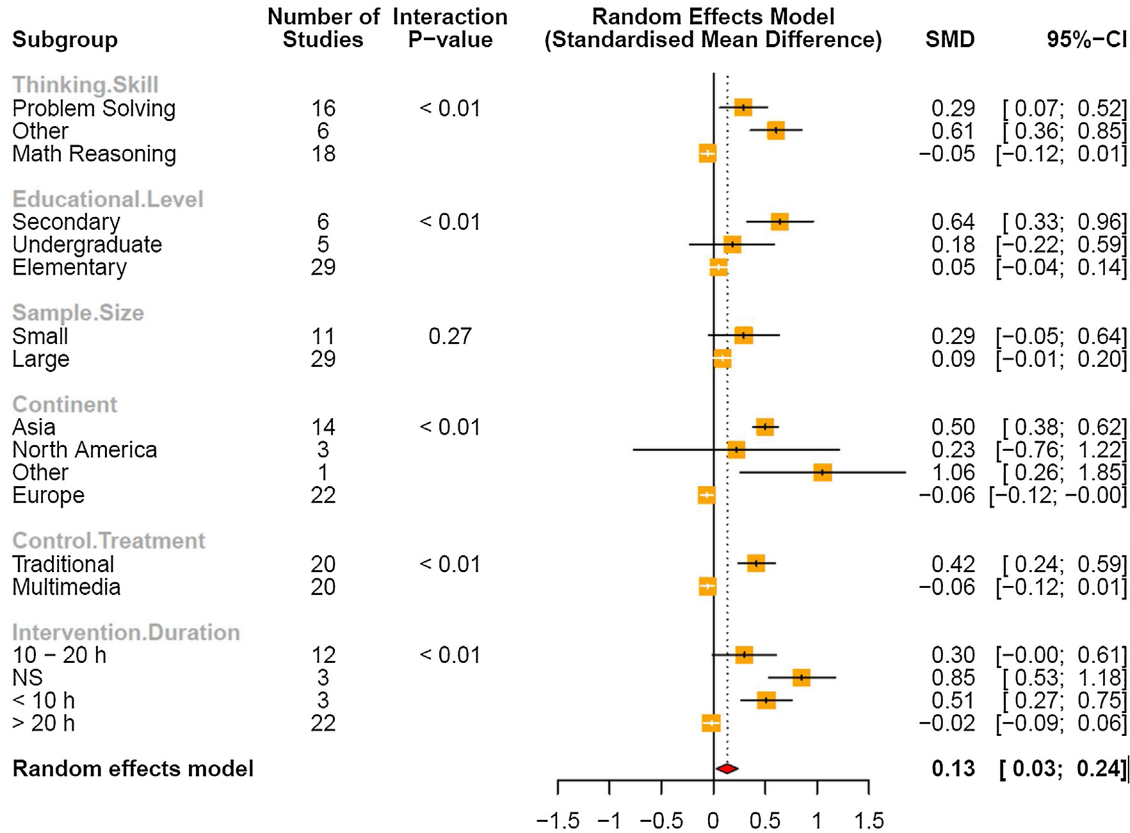

Figure 7. Forest plot of subgroup (moderator variables).

Educational level

In terms of educational level, this study investigates three groups: elementary, secondary, and undergraduate level. The analysis results reveal significant differences in the effect of GBL on students' mathematics HOTS based on the students' level of education ($Qb = 15.09$, $p = .0017$). It indicates that the educational level significantly influence the effectiveness of GBL on students' mathematics HOTS as compared to traditional learning (non-GBL). Among the three measured subject groups, the implementation of GBL to be most effective in the level secondary school ($g = .78$ and 95% CI [.33; .96]. While, two other educational levels (elementary and undergraduate) show the non-significant effect sizes in which their confidence interval include zero (Figure 7).

Sample size

Among the six identified moderator variables, only sample size was found to have no significant impact on the improvement of students' HOTS in GBL. The moderator variable of sample size is categorized into two groups: large and small samples. The findings suggest that there are no significant differences in the impact of GBL on students' mathematics HOTS when considering the sample size of the study ($Qb = 1.22$, $p = .27$) (Figure 7).

Continent

The moderator variable for continent where the study was undertaken is categorized into four groups: Asia, North America, Europe, and other. The findings reveal notable variations in the average effect sizes among these continent groups ($Qb = 76.73$, $p < .001$). The overall $p < .01$ (Figure 7) indicates a significant difference in the effectiveness of GBL across the continents analyzed. Asia shows a positive and significant effect size with $g = .50$. The "Other" category, which includes South America (see Table 1 Appendix), displays an even higher effect size with $g = 1.06$ (95% CI [.26; 1.85]). Although the effect size is substantial and the confidence interval does not include zero, this category includes only one study, so the results should be interpreted with caution. The lack of comparative studies in this region makes it challenging to draw strong conclusions about the effectiveness of GBL in South America. In contrast, North America presents a smaller, statistically non-significant effect size with $g = .23$, indicating a less pronounced impact of GBL in this region. Europe also shows a non-significant effect size with $g = -.06$ (96% CI [-.12; .00]), suggesting that GBL does not exhibit a clear difference compared to conventional learning methods in Europe.

Intervention duration

Treatment duration in the experimental group is categorized into four groups: less than ten hours, 10-20 hours, more than 20 hours, and unspecified. This study revealed significant differences in the impact of GBL on the enhancement of students' mathematics HOTS compared to non-GBL, based on intervention duration ($Qb = 43.44$, $p < .001$). Furthermore, two groups within this intervention duration variable (less than ten hours and unspecified) are suggested to exert an influence. However, the intervention duration category of 10 – 20 hours, the results are not statistically significant because the confidence interval includes 0 (95% CI [-.00; .61]). Additionally, in studies with the longest treatment duration (> 20 hours), GBL was not found to be more effective than non-GBL in improving students' mathematics HOTS (Figure 7).

Control treatment

The control treatment variable is divided into two groups, including traditional and multimedia. The analysis results indicate significant differences in the impact of GBL on students' mathematics HOTS based on the type of control treatment applied ($Qb = 24.01$, $p < .001$). This suggests that the chosen control treatment significantly influences the effectiveness of game-based learning in enhancing students' mathematics HOTS compared to non-game learning. Among the two measured control treatment groups, the implementation of GBL is most effective when compared to traditional non-technological or multimedia-assisted learning ($g = .42$, $p < .001$) (Figure 7).

Discussion

In spite of the increasing trend of incorporating game into the learning process, which has led to numerous empirical research examining the effectiveness of GBL, individual studies yield varied findings regarding its influence on students' mathematics academic thinking skills. Furthermore, a comprehensive synthesis of outcomes from these studies is notably absent. This research seeks to bridge this gap through a meta-analysis approach. The results of this investigation suggest that, on the whole, the implementation of game-based learning positively affects students' mathematics high-order thinking skills.

The current scope of this meta-analysis study focuses effectiveness of GBL on students' mathematics HOTS (Chang et al., 2012). Due to constraints arising from limited data availability, there are challenges in directly comparing the results of this research with other identical studies. However, it is possible to draw comparisons between the results of this study and previous relevant meta-analyses. The findings of this research correspond with the results of numerous earlier meta-analytical studies that have delved into the effectiveness of GBL in the context of general mathematics achievements.

Following the exclusion of outliers, the summary effect analysis reveals an overall positive effect size ($g = .13$), signifying a statistically significant effectiveness of GBL, compared to traditional non-GBL, to enhance students' mathematics HOTS. It is noteworthy, however, that the effect size falls within the 'small effect size' category, as defined by Cohen (2013). This finding notably differs from several earlier meta-analysis studies that reported the large effect or moderate/medium effects of GBL on students' mathematics cognitive performance, such as the findings of Turgut and Temur (2017) and Byun and Joung (2018) which indicate a 'medium/moderate' effect ($g = .792$ and $g = .370$, respectively). However, the results of this study show no substantial difference from the findings observed in study of (Tokac et al., 2019), revealing the 'small but marginally significant' overall effect size ($g = .13$, $p = .02$). The results suggest that, despite multiple meta-analyses consistently indicating the positive influence of GBL on students' cognitive performance in mathematics, the aggregate effect size exhibits variability.

Several prior studies have also explored the impact of GBL on students' HOTS (Harisudin et al., 2019), but with a broader research focus, not limited to the field of mathematics only. In comparison, the results of this study show consistency with those previous studies, indicating positive outcomes. However, the aggregate effect size category found slightly differs from previous meta-analysis study results, such as research examining the effect of GBL on computational thinking abilities, as provided in studies by Lu et al. (2023) and Ma et al. (2023), yielding values of $g = .677$ and $g = .600$, both falling within the 'moderate' category according to Cohen's criteria (Cohen, 2013). In addition, another study examining HOTS in the context of critical thinking reported effect sizes categorized as 'large' effect with $g = .863$ (Mao et al., 2022). These findings reinforce that GBL has a positive effect in enhancing students' HOTS, although the overall effect size remains dynamic.

In addition to information about the aggregate effect size, this meta-analysis study also provides insights into the results of the analysis of moderator variables, including thinking skills measured, educational level, sample size, continent, intervention duration, and control treatment. These findings offer a comprehensive understanding of how various factors contribute to the impact of GBL on students' mathematics HOTS. The exploration of the 'thinking skills measured' variable indicates a significant difference in the average impact size among these groups, suggesting that the effectiveness of GBL on students' HOTS is significantly influenced by the specific measured thinking skills. These findings reinforce previous meta-analyses that reported a positive effect of GBL on students' mathematics HOTS, specifically for certain thinking skills (Aizpurua et al., 2018; Suherman & Vidákovich, 2024), such as critical thinking (Mao et al., 2022) and computational thinking skill (Lu et al., 2023; Ma et al., 2023).

This study examined four educational levels and found significant differences in the impact of GBL on students' mathematics HOTS based on the academic level. These findings align with study of (Turgut & Temur, 2017), indicating a statistically significant difference between groups concerning educational levels, suggesting that the academic level is a contributing factor in providing a positive effect of game-assisted mathematics education on academic achievement (Bakker et al., 2015). In contrast, these findings contradict the results of a previous meta-analysis, which concluded that the impact of GBL on students' cognitive performance remains unaffected by their academic levels (Wang et al., 2022).

Furthermore, it was found that the sample size was the only moderator variable that did not significantly impact the improvement of students' mathematics HOTS in GBL. This finding is consistent with the results of a previous meta-analysis study (Ma et al., 2023), indicating that the sample size factor did not significantly influence the effect of GBL on students' critical thinking. Although not significantly impactful, an interesting aspect of this study's findings is that GBL was considered most effective in enhancing students' mathematics HOTS for small sample sizes compared to large sample sizes. Variations in effect size can be influenced by sample size, with smaller samples typically yielding larger effect sizes compared to larger samples (Slavin & Smith, 2009; Sun et al., 2023). The influence of sample size on effect size variations suggests that GBL tends to show more effective learning outcomes with smaller sample sizes, a trend supported by several previous studies (Petri & Gresse von Wangenheim, 2017).

Another factor that significantly infiuenced the effectiveness of GBL in enhancing students' mathematics HOTS is the continent where the study was conducted (Pires et al., 2019; Sala and Gobet, 2017). Notable variations were observed among continents, with GBL demonstrating its highest effectiveness in the 'other' countries group, particularly in South America, followed by Asia in the second position. This finding has little similarity to a previous meta-analysis study indicating that the use of games in learning, as a learning approach, appears to yield more favorable results in Asian contexts when compared to other continents (Bai et al., 2020). However, this result is still questionable as there is limited empirical support regarding the influence of a country or continent (Tokac et al., 2019) on the impact of GBL on mathematics HOTS. These results need to be further examined in the further investigation.

The type of control treatment also significantly influenced the effectiveness of GBL on students' mathematics HOTS. GBL outperformed traditional non-technological or multimedia-assisted learning, emphasizing the importance of carefully selecting control treatments in experimental designs involving GBL. This finding contradicts the results of a previous study (Wang et al., 2022), that found no statistically significant difference between the control group treated with traditional and multimedia approaches regarding the impact of digital game-based learning on students' cognitive achievement. Limited research has been conducted on this, so further investigation is needed to explore this area.

Furthermore, it was found that the moderator variable of intervention/treatment duration significantly influences the effectiveness of GBL on students' mathematics HOTS. The treatment duration in the experimental group played a crucial role, and interventions with extended durations typically exhibit stronger effects (Korkman et al., 1999). Contradictorily, the findings of this study briefly indicate that GBL proves effective across various durations, but its effectiveness decreases in studies with treatment durations exceeding 20 hours (the longest duration group). This result might seem paradoxical, as one could anticipate that longer interventions would be more effective than shorter ones (Tokac et al., 2019). By disregarding the 'unspecified duration' in this subgroup of moderator variable, the shortest duration (less than ten hours) provided the strongest impact. This is fairly consistent with the results of a previous meta-analysis study (Lu et al., 2023), reporting that GBL interventions between four hours and one week showed the strongest effect on students' critical thinking. This underscores the importance of optimizing intervention duration for optimum impact within an intermediate duration, not too short or too long, to avoid causing boredom in students.

The findings of this meta-analysis suggest that while the overall effect is relatively small, GBL positively impacts students' mathematics HOTS. This indicates that GBL can be a valuable supplementary approach in enhancing mathematics HOTS across different educational levels. However, educators and policymakers should be mindful of specific variables that influence GBL's effectiveness, such as the alignment between the game used and the targeted thinking skills, the educational level of the students, and the duration of the intervention. Given the limitations of the study, including the small number of empirical studies and language restrictions, future research should aim to include a broader range of studies and explore additional moderator variables. This will help to provide a more comprehensive understanding of how GBL can be optimized to support students' mathematical thinking and improve educational practices.

Lastly, our meta-analysis initially revealed a small but statistically significant positive effect. However, upon applying the Trim and Fill method to adjust for potential publication bias, we observed a substantial change in both the magnitude and direction of the effect size. The adjusted effect became slightly negative and statistically non-significant. This marked shift aligns with findings from Carter et al. (2019), who demonstrated that publication bias can lead to overestimation of effect sizes in psychological research. The substantial change we observed suggests that publication bias likely exerted a considerable influence on our original findings, consistent with the meta-meta-analysis by Fanelli et al. (2017) which found that publication bias is prevalent across scientific disciplines. While Egger's test did not indicate significant bias, the notable change in effect size after adjustment cannot be overlooked. This discrepancy between different methods of assessing publication bias echoes the work of Renkewitz and Keiner (2019), who emphasized the importance of using multiple approaches to detect and correct for publication bias. The confidence interval of the adjusted effect now includes zero, further supporting the possibility that the true effect may be negligible or non-existent. This dramatic change from a significant positive effect to a non-significant result raises important questions about the robustness and balance of the available literature in this field, reminiscent of the "decline effect" described by Schooler (2011) in psychological science. Our findings suggest that published studies may not fully represent the entire landscape of conducted research, with a potential bias towards reporting positive or significant outcomes, a phenomenon well-documented by Ioannidis (2005). They also serve as a reminder to interpret meta-analytic results with caution, especially in fields where publication bias may be prevalent, as highlighted by Ferguson and Heene (2012) in their critique of psychological meta-analyses.

Limitation and implication

Despite offering valuable insights into GBL and its impact on students' mathematics HOTS, this study has several limitations. The review is constrained by a small number of empirical studies—only thirteen were included—which may affect the generalizability of the findings to broader populations and diverse educational contexts. The limited scope is partly due to difficulties accessing the Clarivate Web of Science (WoS) database and relevant theses or dissertations. Additionally, the study was restricted to English-language articles due to language limitations. Furthermore, although six moderator variables were analyzed, other influential factors may not have been fully considered, suggesting the need for more comprehensive analyses in future research. Additionally, this analysis addresses each moderator individually, and does not account for potential interactions between moderators, which may collectively influence the effectiveness of GBL.

Acknowledging these limitations is crucial for a deeper understanding and for guiding improvements in future GBL and mathematics education research. This underscores the necessity for broader investigations involving a larger number of empirical studies across diverse populations and educational settings to provide a more complete understanding of the impact of GBL on mathematics HOTS. Additionally, future studies could explore other potential moderator variables not covered in this meta-analysis, such as the specific characteristics of GBL interventions or students' prior knowledge levels. Furthermore, future research could explore the combined effects of multiple moderators through more complex analytical approaches, such as interaction effects in meta-regression models. Adopting this broader approach will contribute to a more comprehensive understanding of GBL dynamics in mathematics education.

A notable limitation of this study also is the potential for publication bias, as evidenced by the significant reduction in effect size following the Trim and Fill adjustment. While Egger's test did not indicate substantial bias, the findings from the Trim and Fill method suggest that missing studies may have skewed the original results, possibly inflating the effect size. This highlights the need for caution when interpreting the findings, as the dataset may not be as robust as initially assumed. Additionally, the analysis is constrained by the data available, with potential variability in study design and sample characteristics across the included studies contributing to the observed bias. Relying solely on statistical corrections like Trim and Fill may not fully account for all dimensions of publication bias, leaving room for further exploration. Future research should prioritize replicating these findings with more comprehensive datasets and employ a broader range of methods for detecting and correcting publication bias. Expanding the scope of databases searched, encouraging pre-registration of studies, and considering additional tools such as p-curve analysis or selection models could provide deeper insights into the influence of publication bias, ultimately enhancing the reliability and validity of meta-analytic outcomes.

The implications of this meta-analysis for mathematics education are significant. Despite the relatively small effect size, GBL has shown potential as a supplementary tool for enhancing students' mathematics HOTS. Educators should consider integrating GBL strategies into their curricula to stimulate students' mathematics thinking. However, it is essential to carefully select and design GBL interventions to align with specific educational goals, the types of thinking skills targeted, and the educational level. Future research should also investigate how game characteristics and the interplay between students' prior knowledge and GBL can provide insights into customizing interventions to better meet individual learning needs, particularly in enhancing their mathematics HOTS. By addressing these aspects, future studies can provide a more comprehensive understanding of how GBL can be effectively utilized to support and enhance learning quality in mathematics classroom.

Conclusion

Considering the varied conclusions derived from individual studies assessing the effectiveness of GBL on students' mathematics HOTS, this meta-analysis aims to offer an updated synthesis of the impact of GBL on students' mathematics HOTS. It concludes that GBL positively affects students' mathematics HOTS, with an overall effect size falling within the 'small' category. Albeit the aggregated effect size found is relatively small, it is sufficient to provide insights that GBL can be considered as an alternative approach to enhance students' mathematics HOTS across various educational levels. This study also reveals that variables such as the type of thinking skills measured, educational level, continent where the study was conducted, type of control treatment, and intervention duration tend to significantly influence the impact of GBL on students' mathematics HOTS. While the sample size factor showed no significant impact. However, the factor of sample size showed no significant impact on the effectiveness of GBL. Overall, these findings indicate that while GBL can support the development of students' mathematics HOTS, attention should be given to the specific variables affecting its effectiveness.

CRediT authorship contribution statement

Bambang Sri Anggoro: supervision, funding acquisition, writing – review & editing.

Andi Harpeni Dewantara: methodology, writing - original draft, formal analysis.

Suherman Suherman: conceptualization, writing - original draft, formal analysis, methodology, editing, and visualization.

Rosida Rakhmawati Muhammad: writing - original draft, formal analysis.

Sari Saraswati: writing – review & editing.

Acknowledgements

The authors gratefully acknowledge to the Universitas Islam Negeri Raden Intan Lampung, Indonesia, for their research support. We also extend our thanks for Doctoral School of Education at the University of Szeged, Hungary, Hasyim Asy'ari University, Indonesia, and Indonesian Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan - LPDP) of the Indonesian Ministry, for their valuable supports, suggestions and discussions that contributed to this research.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.psicoe.2024.500158.

[1] References labelled with an * were included in the meta-analysis.

References[1]

Acharya, B. R. (2017). Factors affecting difficulties in learning mathematics by mathematics learners. *International Journal of Elementary Education*, *6*(2), 8-8. https://doi.org/10.11648/j.ijeedu.20170602.11

Aizpurua, A., Lizaso, I., & Iturbe, I. (2018). Estrategias de aprendizaje y habilidades de razonamiento de estudiantes universitarios. *Revista de Psicodidáctica*, *23*(2), 110–116. https://doi.org/10.1016/j.psicod.2018.01.001

Al-Absi*, M. (2017). The effect of using puzzles and games on students' mathematical thinking at the faculty of educational sciences and arts (UNRWA). *An-Najah University Journal for Research - B (Humanities)*, *31*(10), 1867–1888. https://doi.org/10.35552/0247-031-010-007

Angelelli, C. V., Ribeiro, GM de C, Severino, M. R., Johnstone, E., Borzenkova, G., & da Silva, D. C. O. (2023). Developing critical thinking skills through gamification. *Thinking Skills and Creativity*, *49*, 101354. https://doi.org/10.1016/j.tsc.2023.101354

Asri, A. S. M., & Jamaludin, K. A. (2022). Potential scratch games in developing students' thinking skills. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, *7*(12), e002004. https://doi.org/10.47405/mjssh.v7i12.2004

Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review*, *30*, 100322. https://doi.org/10.1016/j.edurev.2020.100322

Bakker*, M., van den Heuvel-Panhuizen, M., & Robitzsch, A. (2015). Effects of playing mathematics computer games on primary school students' multiplicative reasoning ability. *Contemporary Educational Psychology*, *40*, 55–71. https://doi.org/10.1016/j.cedpsych.2014.09.001

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: a practical tutorial. *Evidence Based Mental Health*, *22*(4), 153–160. https://doi.org/10.1136/ebmental-2019-300117

Barz, N., Benick, M., Dörrenbächer-Ulrich, L., & Perels, F. (2024). 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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Bourke, B. (2019). Using gamification to engage higher-order thinking skills. In J. Keengwe, & R. Byamukama (Eds.), *Handbook of research on promoting higher-order skills and global competencies in life and work* (pp. 632–652). IGI Global. https://doi.org/10.4018/978-1-7998-3022-1.ch033

Byun, J., & Joung, E. (2018). Digital game-based learning for K–12 mathematics education: A meta-analysis. *School Science and Mathematics*, *118*(3–4), 113–126. https://doi.org/10.1111/ssm.12271

Card, N. A. (2011). *Applied meta-analysis for social science research*. Guilford Publications.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Chang*, K.-E., Wu, L.-J., Weng, S.-E., & Sung, Y.-T. (2012). Embedding game-based problem-solving phase into problem-posing system for mathematics learning. *Computers & Education*, *58*(2), 775–786. https://doi.org/10.1016/j.compedu.2011.10.002

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach*. Sage Publication.

Demirel*, T., & Yilmaz, T. K. (2019). The effects of mind games in math and grammar courses on the achievements and perceived problem-solving skills of secondary school students. *British Journal of Educational Technology*, *50*(3), 1482–1494. https://doi.org/10.1111/bjet.12624

Deng, R., Benckendorff, P., & Gannaway, D. (2020). Learner engagement in MOOCs: Scale development and validation. *British Journal of Educational Technology*, *51*(1), 245–262. https://doi.org/10.1111/bjet.12810

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Von Elm, E., & Gamble, C. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS One*, *3*(8), Article e3081. https://doi.org/10.1371/journal.pone.0003081

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Emihovich*, B., Roque, N., & Mason, J. (2022). Can video gameplay improve undergraduates' problem-solving skills? *International Journal of Game-Based Learning*, *10*(2), 1–18. https://doi.org/10.4018/IJGBL.2020040102

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, *114*(14), 3714–3719. https://doi.org/10.1073/pnas.1618569114

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/17456916124590

Freitas, S. A. A., Lacerda, A. R. T., Calado, P. M. R. O., Lima, T. S., & Canedo, E. D. (2017). *Gamification in education: A methodology to identify student's profile*. pp. 1–8.

Ganapathy, M., Mehar Singh, M. K., Kaur, S., & Kit, L. W. (2017). Promoting higher order thinking skills via teaching practices. *3L The Southeast Asian Journal of English Language Studies*, *23*(1), 75–85. https://doi.org/10.17576/3L-2017-2301-06

Harisudin*, L., Susanto, S., & Hobri, H. (2019). The development of mathematics learning tools through the bridge games based on lesson study for learning community and its relationship with the higherorder thinking skills in probability theory. *Journal of Physics: Conference Series*, *1211*, 012075. https://doi.org/10.1088/1742-6596/1211/1/012075

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing meta-analysis with R: A hands-on guide*. CRC press.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series c (Applied Statistics)*, *28*(1), 100–108.

Hartt, M., Hosseini, H., & Mostafapour, M. (2020). Game on: Exploring the effectiveness of game-based learning. *Planning Practice & Research*, *35*(5), 589–604. https://doi.org/10.1080/02697459.2020.1778859

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *Bmj*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Ho, G.-Y., Leonhard, M., Volk, G. F., Foerster, G., Pototschnig, C., Klinge, K., Granitzka, T., Zienau, A.-K., & Schneider-Stickler, B. (2019). Inter-rater reliability of seven neurolaryngologists in laryngeal EMG signal interpretation. *European Archives of Oto-Rhino-Laryngology*, *276*(10), 2849–2856. https://doi.org/10.1007/s00405-019-05553-y

Hsu, C.-Y., Liang, J.-C., & Tsai, M.-J. (2020). Probing the structural relationships between teachers' beliefs about game-based teaching and their perceptions of technological pedagogical and content knowledge of games. *Technology, Pedagogy and Education*, *29*(3), 297–309. https://doi.org/10.1080/1475939X.2020.1752296

Indriani*, M. N., Isnarto, I., & Mariani, S. (2019). The implementation of PBL (problem based learning) model assisted by monopoly game media in improving critical thinking ability and self confidence. *Journal of Primary Education*, *8*(2), 200–208. https://doi.org/10.15294/jpe.v8i2.25991

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.1004085

Ioannidis, J. P., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543–549. https://doi.org/10.1016/j.jclinepi.2004.10.019

Ioannidis, J. P., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Cmaj*, *176*(8), 1091–1096. https://doi.org/10.1503/cmaj.060410

Izzati, G. N., Dwijanto, D., & Cahyono, A. N. (2022). The effectivity of problem-based learning assisted by snakes and ladders games on students divergent thinking ability. *Jurnal Pijar Mipa*, *17*(6), 732–736. https://doi.org/10.29303/jpm.v17i6.4177

Jabbar, A. I. A., & Felicia, P. (2015). Gameplay engagement and learning in game-based learning. *Review of Educational Research*, *85*(4), 740–779. https://doi.org/10.3102/0034654315577210

Ke*, F. (2019). Mathematical problem solving and learning in an architecture-themed epistemic game. *Educational Technology Research and Development*, *67*(5), 1085–1104. https://doi.org/10.1007/s11423-018-09643-2

Korkman, M., Barron-Linnankoski, S., & Lahti-Nuuttila, P. (1999). Effects of age and duration of reading instruction on the development of phonological awareness, rapid naming, and verbal memory span. *Developmental Neuropsychology*, *16*(3), 415–431. https://doi.org/10.1207/S15326942DN1603_24

Lee, T. Y., Mauriello, M. L., Ahn, J., & Bederson, B. B. (2014). CTArcade: Computational thinking with games in school age children. *International Journal of Child-Computer Interaction*, *2*(1), 26–33. https://doi.org/10.1016/j.ijcci.2014.06.003

Li, X., Dusseldorp, E., Su, X., & Meulman, J. J. (2020). Multiple moderator meta-analysis using the R-package Meta-CART. *Behavior Research Methods*, *52*(6), 2657–2673. https://doi.org/10.3758/s13428-020-01360-0

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE publications, Inc.

López-Díaz, M. T., & Peña, M. (2021). Mathematics training in engineering degrees: An intervention from teaching staff to students. *Mathematics*, *9*(13), 1475-1475. https://doi.org/10.3390/math9131475

Lovianova, I. V., Yu Kaluhin, R., Kovalenko, D. A., Rovenska, O. G., & Krasnoshchok, A. V. (2022). Development of logical thinking of high school students through a problem-based approach to teaching mathematics. *Journal of Physics: Conference Series*, *2288*(1), 012021. https://doi.org/10.1088/1742-6596/2288/1/012021

Lu, Z., Chiu, M. M., Cui, Y., Mao, W., & Lei, H. (2023). Effects of game-based learning on students' computational thinking: A meta-analysis. *Journal of Educational Computing Research*, *61*(1), 235–256. https://doi.org/10.1177/07356331221100740

Ma, J., Zhang, Y., Zhu, Z., Zhao, S., & Wang, Q. (2023). Game-based learning for students' computational thinking: A meta-analysis. *Journal of Educational Computing Research*, *61*(7), 1430–1463. https://doi.org/10.1177/07356331231178948

Mao, W., Cui, Y., Chiu, M. M., & Lei, H. (2022). Effects of game-based learning on students' critical thinking: A meta-analysis. *Journal of Educational Computing Research*, *59*(8), 1682–1708. https://doi.org/10.1177/07356331211007098

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education*, *70*, 29–40. https://doi.org/10.1016/j.compedu.2013.07.033

Mumcu, H. Y. (2018). Examining mathematics department students' views on the use of mathematics in daily life. *International Online Journal of Education and Teaching (IOJET)*, *5*(1), 61–80.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Tricco, A. C., Welch, V., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 1–8. https://doi.org/10.1136/bmj.n71

Petri, G., & Gresse von Wangenheim, C. (2017). How games for computing education are evaluated? A systematic literature review. *Computers & Education*, *107*, 68–90. https://doi.org/10.1016/j.compedu.2017.01.004

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46. https://doi.org/10.3102/0034654319877153

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*(8), 424–432. https://doi.org/10.3102/0013189X13507104

Pires, F., Lima, F. M. M. D., Melo, R., Bernardo, J. R. S., & Freitas, R. d. (2019). Gamification and engagement: development of computational thinking and the implications in mathematical learning. https://doi.org/10.1109/icalt.2019.00112

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, *50*(4), 258–283. https://doi.org/10.1080/00461520.2015.1122533

Posit Team. (2023). *RStudio: Integrated development environment for R (2023.3.1.446)*.

Prieto-Rodriguez, E., Sincock, K., & Blackmore, K. (2020). STEM initiatives matter: Results from a systematic review of secondary school interventions for girls. *International Journal of Science Education*, *42*(7), 1144–1161. https://doi.org/10.1080/09500693.2020.1749909

Qian, M., & Clark, K. R. (2016). Game-based Learning and 21st century skills: A review of recent research. *Computers in Human Behavior*, *63*, 50–58. https://doi.org/10.1016/j.chb.2016.05.023

Renkewitz, F., & Keiner, M. (2019). How to Detect Publication Bias in Psychological Research: A Comparative Evaluation of Six Statistical Methods. *Zeitschrift Für Psychologie*, *227*(4), 261–279. https://doi.org/10.1027/2151-2604/a000386

Riahi, S. M., & Mokhayeri, Y. (2017). Methodological issues in a meta-analysis. *Current Medical Research and Opinion*, *33*(10), 1813-1813. https://doi.org/10.1080/03007995.2017.1359152

Rothstein, H. R. (2005). *Publication bias in meta-analysis. Publication Bias in Meta-Analysis-Prevention, Assessment and Adjustment*. John Wiley & Sons. https://doi.org/10.1002/0470870168

Sala*, G., & Gobet, F. (2017). Does chess instruction improve mathematical problem-solving ability? Two experimental studies with an active control group. *Learning & Behavior*, *45*(4), 414–421. https://doi.org/10.3758/s13420-017-0280-3

Scherer, R., & Beckmann, J. F. (2014). The acquisition of problem solving competence: Evidence from 41 countries that math and science education matters. *Large-Scale Assessments in Education*, *2*(1), 10. https://doi.org/10.1186/s40536-014-0010-7

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437-437. https://doi.org/10.1038/470437a

Shu, L., & Liu, M. (2019). Student engagement in game-based learning: A literature review from 2008 to 2018. *Journal of Educational Multimedia and Hypermedia*, *28*(2), 193–215.

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500–506. https://doi.org/10.3102/0162373709352369

Smith, K., Maynard, N., Berry, A., Stephenson, T., Spiteri, T., Corrigan, D., Mansfield, J., Ellerton, P., & Smith, T. (2022). Principles of problem-based learning (PBL) in STEM education: Using expert wisdom and research to frame educational practice. *Education Sciences*, *12*(10), 728. https://doi.org/10.3390/educsci12100728

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., & Schmid, C. H. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *Bmj*, *343*. https://doi.org/10.1136/bmj.d4002

Suherman, S., Prananda, M. R., Proboningrum, D. I., Pratama, E. R., Laksono, P., & Amiruddin, A. (2020). Improving higher order thinking skills (hots) with project based learning (pjbl) model assisted by geogebra. *Journal of Physics: Conference Series*, *1467*(1), Article 012027. https://doi.org/10.1088/1742-6596/1467/1/012027

Suherman, S., & Vidákovich, T. (2022). Assessment of mathematical creative thinking: A systematic review. *Thinking Skills and Creativity*, *44*, Article 101019. https://doi.org/10.1016/j.tsc.2022.101019

Suherman, S., & Vidákovich, T. (2024). Relationship between ethnic identity, attitude, and mathematical creative thinking among secondary school students. *Thinking Skills and Creativity*, *51*, Article 101448. https://doi.org/10.1016/j.tsc.2023.101448

Sun, L., Guo, Z., & Hu, L. (2023). Educational games promote the development of students' computational thinking: A meta-analytic review. *Interactive Learning Environments*, *31*(6), 3476–3490. https://doi.org/10.1080/10494820.2021.1931891

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *Bmj*, *320*(7249), 1574–1577. https://doi.org/10.1136/bmj.320.7249.1574

Taja-on*, E. P. (2019). Game-aided instruction: Enhancing critical thinking through logical-mathematical games. *Online Submission*, *1*(1), 1–15. https://doi.org/10.13140/RG.2.2.16436.17288

Tokac, U., Novak, E., & Thompson, C. G. (2019). Effects of game-based learning on students' mathematics achievement: A meta-analysis. *Journal of Computer Assisted Learning*, *35*(3), 407–420. https://doi.org/10.1111/jcal.12347

Turgut, S., & Temur, Ö. D. (2017). The effect of game-assisted mathematics education on academic achievement in Turkey: A meta-analysis study. *International Electronic Journal of Elementary Education*, *10*(2), 195–206. https://doi.org/10.26822/iejee.2017236115

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Wang, L.-H., Chen, B., Hwang, G.-J., Guan, J.-Q., & Wang, Y.-Q. (2022). Effects of digital game-based STEM education on students' learning achievement: A meta-analysis. *International Journal of STEM Education*, *9*(1), 1–13. https://doi.org/10.1186/s40594-022-00344-0

White, K., & McCoy, L. P. (2019). Effects of game-based learning on attitude and achievement in elementary mathematics. *Networks: An Online Journal for Teacher Research*, *21*(1), 1–17. https://doi.org/10.4148/2470-6353.1259

Wibawa, K. A., Payadnya, P. A. A., Yasa, G. U., & Prahmana, R. C. I. (2022). The Learning Trajectory of Entrepreneurship Arithmetic Content using a Traditional Market. *Mathematics Teaching Research Journal*, *14*(3), 144–169.

Yeh*, C. Y. C., Cheng, H. N. H., Chen, Z.-H., Liao, C. C. Y., & Chan, T.-W. (2019). Enhancing achievement and interest in mathematics learning through Math-Island. *Research and Practice in Technology Enhanced Learning*, *14*(1), 5-5. https://doi.org/10.1186/s41039-019-0100-9

Yerizon, Y., Triwani, T., & Musdi, E. (2022). Effectiveness of mathematics learning devices based on flipped classroom to Improve mathematical critical thinking ability students. *International Journal of Education and Management Engineering*, *12*(3), 41–46. https://doi.org/10.5815/ijeme.2022.03.05