Does Game-Based Learning Improve K–12 Mathematics Achievement Beyond Equivalent Practice? A Systematic Review and Multilevel Meta-Analysis of Comparator and Assessment Conditions

Qiang Zhang, Mohamad Ikram Zakaria, Norulhuda Ismail

Universiti Teknologi Malaysia, Johor, Bahru Johor, Malaysia

DOI: https://doi.org/10.47772/IJRISS.2026.100701137

Received: 09 August 2026; Accepted: 14 August 2026; Published: 22 August 2026

ABSTRACT

Game-based learning (GBL) is widely used in mathematics education, yet its added value depends on what comparison groups receive and how learning is assessed. This systematic review and multilevel meta-analysis examined whether complete digital or non-digital games improve objective K–12 mathematics achievement relative to concurrent non-game instruction or practice, with attention to comparator equivalence and assessment proximity. Four database searches yielded 2,633 records. Independent verification of Rayyan's deduplication history confirmed that 859 duplicates were removed, leaving 1,774 records for title–abstract screening; 150 candidate reports were then sought for retrieval. Sixty reports were not retrieved. Eighty eligible reports represented 79 independent study families. The primary three-level random-effects model included 42 dependent effects from 19 families and used cluster-robust CR2 inference. GBL produced a positive average effect, $g = 0.368$, 95% CI [0.176, 0.560], $p < .001$, with substantial residual heterogeneity (approximate $I^2 = 65.5\%$). Neither comparator nor assessment conditions, nor the other prespecified moderators, reached statistical significance; the limited data did not support conclusions of equivalence. The pooled estimate remained positive across working-correlation assumptions, randomized-design restrictions, the expanded synthesis set, and leave-one-family-out analyses. However, no randomized family was judged at low risk of bias, most non-randomized families were at serious or critical risk, and exploratory diagnostics suggested possible small-study effects. On average, complete mathematics GBL was associated with better achievement than the available non-game comparators, but its added value over fully equivalent practice remains uncertain. Stronger randomized value-added trials using matched practice and both proximal and independent distal assessments are needed.

Keywords: game-based learning; mathematics education; K–12; multilevel meta-analysis; comparator equivalence; assessment proximity

INTRODUCTION

Objective and rationale

Mathematics education is expected to develop more than the accurate execution of procedures. Students must also build conceptual understanding, use representations, reason strategically, solve unfamiliar problems, and apply knowledge across contexts (National Council of Teachers of Mathematics [NCTM], 2014). Game-based learning (GBL) is often proposed as a way to organize these demands into purposeful activity. A mathematics game can require learners to make repeated numerical decisions, test strategies, respond to feedback, and revise their actions in pursuit of a goal. The same general structure can support arithmetic fluency in early grades, strategic reasoning about fractions or algebra, spatial and geometric thinking, or multi-step problem solving. In each case, the instructional promise is that mathematical activity is not merely presented alongside play but becomes part of what players must understand and do to progress.

This promise is theoretically plausible because games can create recurring cycles of goals, player actions, feedback, and renewed engagement (Garris et al., 2002). Game environments can also combine cognitive, motivational, affective, and sociocultural supports: they can focus attention on a task, provide immediate information about consequences, encourage persistence after error, and situate learning in individual or collaborative activity (Plass et al., 2015). When mathematical content is intrinsically integrated with the core mechanics, succeeding in the game requires learners to use the intended mathematics rather than complete conventional exercises merely to unlock a separate reward (Habgood & Ainsworth, 2011). These features make complete games a potentially powerful environment for sustained mathematical practice and sense making.

However, the presence of a game does not guarantee better learning. Learners may attend to points, competition, narrative, or visual effects that are only weakly connected to the mathematical objective. A game may increase time on task without improving the quality of practice, or it may bundle mathematics with additional technology, teacher attention, feedback, or instructional time. Conversely, a carefully designed non-game practice condition may provide the same content, dosage, feedback, and support without the game structure. The central question is whether complete GBL improves objective mathematics achievement relative to the instruction or practice students would otherwise receive, and whether the answer changes with comparator and assessment conditions.

Earlier reviews generally suggest that games can support mathematics learning, but their scopes and construct definitions differ. Some syntheses focus on digital games, some mix complete games with gamified learning environments, some examine broad school subjects or serious games, and others restrict the population, mathematical outcome, or publication period. Their pooled effects do not isolate the incremental value of the game format over equivalent practice. The present synthesis therefore examines the counterfactual, the mathematics outcome being measured, and the dependence among multiple outcomes reported by the same study.

Complete game-based learning in K–12 mathematics education

In this review, a complete game is an activity with identifiable rules, a goal or end condition, player actions or decisions, challenge or progression, and feedback or consequences that together form a coherent game loop. Complete GBL occurs when learners participate in such a game in order to develop mathematical knowledge or skill. Throughout this review, GBL refers specifically to learning through a complete digital or non-digital game. Eligible forms can therefore include computer, tablet, or mobile games as well as board, card, dice, or other physical games, provided that the mathematical activity is embedded in a coherent system of play rather than appended as an unrelated exercise.

This definition separates GBL from several adjacent constructs. Gamification is the use of game design elements in a context that remains fundamentally non-game, such as adding points, badges, levels, or leader boards to conventional exercises (Deterding et al., 2011; Sailer & Homner, 2020). Gamification may influence learning, but it does not test the same instructional object as participation in a complete game. Serious games, by contrast, are complete games designed primarily for purposes beyond entertainment and can be a form of GBL when their central purpose is mathematics learning (Wouters et al., 2013). Digital game-based learning is a medium-specific subset of GBL, not a synonym for the whole construct. Likewise, ordinary computer-assisted instruction, quiz platforms, dynamic visualizations, simulations without a game loop, and narrative "skins" around routine worksheets are not automatically GBL.

The conceptual boundary matters because the mechanisms attributed to a complete game are systemic. In the game-cycle model, player judgments lead to actions, actions produce feedback, and feedback informs the next attempt (Garris et al., 2002). In mathematics, this cycle may support repeated retrieval, procedural refinement, hypothesis testing, or strategic comparison. Plass et al. (2015) further argue that learning with games cannot be reduced to motivation alone; game environments coordinate cognitive processing, affect, motivation, and social participation. Habgood and Ainsworth's (2011) principle of intrinsic integration sharpens this claim: learning should be connected to the core mechanics and most engaging parts of play. If the mathematics is

incidental to winning, observed engagement may not translate into mathematical achievement. The intervention of interest is therefore a complete game system in which mathematical activity is functionally connected to player progress.

Complete games can instantiate this connection in different ways. A practice-oriented game may use repeated, progressively difficult arithmetic decisions and immediate corrective feedback. A strategy game may require players to compare quantities, plan transformations, allocate resources, or infer spatial relations. Collaborative and competitive games may externalize reasoning through explanation, negotiation, and shared decision making. Yet these forms are heterogeneous, and their effectiveness should not be inferred from the label "game" alone. The alignment among game mechanics, mathematical content, learners, classroom implementation, and assessment remains decisive.

Mathematics game-based learning interventions

We use mathematics GBL intervention to refer to a school, classroom, or structured learning environment in which K–12 learners participate in complete gameplay intended to improve objective mathematics learning. The defining comparison is between that environment and a concurrent condition in which students receive non-game instruction or practice. This contrast is narrower than asking whether games are associated with learning and broader than evaluating one platform: it covers digital and non-digital games, multiple grade bands, and objective outcomes ranging from numeracy and arithmetic to algebra, geometry, mathematical reasoning, problem solving, retention, and transfer.

The closest quantitative syntheses provide a favorable but incomplete evidence base. Byun and Joung (2018) meta-analyzed digital game-based learning for K–12 mathematics, whereas Tokac et al. (2019) synthesized effects of GBL on mathematics achievement across a broader pre-K–12 range. Conmy (2023) included digital, mobile, and non-digital games but restricted the inquiry to K–12 studies conducted in the United States. Kaçmaz and Dubé (2022) examined the pedagogical approaches embodied in mathematics games and the factual, procedural, and conceptual knowledge they targeted. Together, these reviews show that game design, medium, pedagogy, and knowledge type vary substantially across the mathematics-game literature; they also suggest that positive average effects cannot be interpreted as a single, uniform property of all games.

Recent reviews extend the field in other directions. Ersen and Ergül (2022) and Hui and Mahmud (2023) described trends and cognitive or affective outcomes in recent mathematics GBL research. Barz et al. (2024) synthesized digital GBL across school subjects and across cognitive, metacognitive, and affective-motivational outcomes. Anggoro et al. (2025) focused specifically on higher-order mathematical thinking skills. At an umbrella level, Yao (2026) synthesized 20 meta-analyses under the broad label of gamified mathematics learning and found that research quality affected the credibility of estimated benefits. These contributions strengthen the case that games may support learning, but their broad or specialized scopes do not permit primary-study recoding of both comparator equivalence and assessment proximity within a construct-strict K–12 mathematics evidence base.

Existing syntheses also differ in whether they include gamification, serious games, digital-only interventions, non-digital games, affective outcomes, higher education, or studies without a clearly comparable non-game condition. Consequently, their pooled estimates answer related but non-identical questions. The present review brings the relevant design features together by restricting the intervention to complete games, including digital and non-digital K–12 mathematics interventions, requiring a concurrent non-game comparator, coding comparator equivalence and assessment condition, and retaining dependent effects in a multilevel model. This design helps distinguish advantages associated with game structure from those associated with weaker comparators or closely aligned assessments.

Moderators of learning mathematics with games

Beyond the mean effect, mathematics GBL studies vary in assignment design, comparator strength, game medium and pedagogy, learner age, mathematical content, implementation, outcome domain, assessment

proximity, and timing. These differences can change both the magnitude and the meaning of an effect size. We therefore organized potential moderators into study and comparator characteristics, intervention and learner characteristics, and outcome and assessment characteristics. Comparator and assessment conditions were given priority because they determine, respectively, what instructional contrast is being estimated and how closely the measured performance is tied to the intervention.

Study and comparator characteristics

Assignment design can influence estimated intervention effects. Quasi-experimental studies may compare pre-existing groups that differ in teacher interest, implementation readiness, prior achievement, or access to resources, whereas randomized studies reduce—although they do not eliminate—these sources of confounding. Across education reviews, quasi-experimental designs have tended to produce larger effects than randomized designs, and other methodological features such as small scale and experimenter-developed measures have also been associated with larger estimates (Cheung & Slavin, 2016). Design must therefore be examined as a characteristic of the evidence, not treated as interchangeable background information.

The comparator is equally important because every intervention effect is relative to a counterfactual. A game compared with business-as-usual instruction or a passive condition may differ not only in game structure but also in content, time on task, amount of practice, feedback, technology access, and teacher attention. An active but not fully equivalent comparator controls some of these features, whereas equivalent active practice attempts to match them more closely. As the counterfactual becomes stronger, the contrast increasingly approaches the incremental contribution of organizing the same mathematical practice as a game. Effect-size interpretation in education therefore requires attention to both the treatment and what the comparison group actually received (Cheung & Slavin, 2016; Kraft, 2020).

For this review, comparators were classified as business-as-usual/passive, active but not fully equivalent, or equivalent active practice. Equivalence was also represented by explicit matching on mathematical content, instructional time or practice dosage, technology access, and teacher attention or support. This structure addresses the title question directly. A positive effect against a weak comparator shows that a GBL package outperformed that alternative; it does not, by itself, prove that the game format added value beyond equivalent practice. Conversely, evidence from well-matched active comparisons provides a more stringent test, even when the resulting effect is smaller.

Intervention and learner characteristics

Game interventions differ in medium, pedagogical orientation, and mathematical demand. Digital games can automate feedback, adapt difficulty, and present dynamic or individualized representations, whereas non-digital games can make rules and quantities physically manipulable and can support face-to-face explanation and negotiation. Neither format is inherently a more complete game, and technology alone should not be mistaken for the active ingredient. Reviews of educational games show that direct instruction, practice, discovery, experiential, and constructivist approaches coexist and target different types of mathematical knowledge (Kaçmaz & Dubé, 2022). The effectiveness of GBL may therefore depend on whether the core mechanics are aligned with numeracy, arithmetic, algebra, fractions, geometry, spatial reasoning, or broader problem-solving demands rather than on digitality alone.

Learners and instructional contexts also vary across K–12. Games suitable for Kindergarten or the early grades may emphasize counting, magnitude, and basic operations with substantial guidance; games for older learners may require symbolic reasoning, multi-step strategy, or self-regulation. Appropriately designed games may support multiple age groups, but developmental level, prior knowledge, classroom norms, and teacher facilitation can shape how students interpret goals and feedback (Plass et al., 2015). Because these learner characteristics were not consistently reported across primary studies, the present synthesis treated grade band as the principal modelable learner/context characteristic and did not impose directional hypotheses for grade, digitality, mathematics domain, or publication year.

Outcome and assessment characteristics

The same intervention can appear more or less effective depending on what is measured. A proximal assessment samples the content and procedures practiced in the game and is often administered immediately after the intervention. A distal, transfer, or standardized assessment requires learners to apply knowledge beyond the specific game tasks or uses an instrument developed independently of the intervention. Delayed assessments examine retention after direct gameplay has ended. These distinctions overlap but are not identical: a standardized test is often distal, yet standardization alone does not fully describe content alignment or timing.

Proximal and researcher-developed measures may be more sensitive to the exact knowledge taught, but they can also yield larger estimates because of treatment–outcome alignment. In a cross-review analysis, experimenter-made measures produced substantially larger effects than independent measures (Cheung & Slavin, 2016). Mathematics technology syntheses have likewise reported differences between standardized and nonstandardized outcomes (Li & Ma, 2010). Distal and standardized measures provide a more demanding test of generalization, whereas delayed measures test persistence. Because intervention effect sizes depend on the outcome scale and the counterfactual against which change is measured (Kraft, 2020), assessment condition was treated as a central feature of effect interpretation.

The present study

The present systematic review and multilevel meta-analysis synthesized comparative evidence on complete digital and non-digital GBL for objective K–12 mathematics outcomes. Eligible studies compared a complete mathematics game with concurrent non-game instruction or practice using an individually randomized, cluster-randomized, or quasi-experimental design. Reports, independent study families, effect sizes, and synthesis-specific k values were tracked separately. By modeling multiple dependent effects within study families, the review retained outcome and time-point information that would be lost by selecting a single estimate while avoiding the assumption that all effects were independent.

Accordingly, the review addressed five questions: (RQ1) What is the overall effect of complete GBL on objective K–12 mathematics achievement? (RQ2) Does the effect vary by comparator condition or comparator-equivalence score? (RQ3) Does it vary by assessment condition? (RQ4) Do study design, digitality, grade band, mathematics domain, or publication year moderate the effect? (RQ5) How robust are the findings to dependence assumptions, synthesis set, design and risk-of-bias restrictions, influential families, and small-study-effect diagnostics?

We expected a positive overall effect, with larger estimates for quasi-experimental designs, weaker comparators, and proximal assessments. No directional hypotheses were specified for digitality, grade band, mathematics domain, or publication year.

METHOD

This systematic review and meta-analysis was designed and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement and the PRISMA-S extension for reporting literature searches (Page et al., 2021; Rethlefsen et al., 2021). The review question was structured using the Population, Intervention, Comparator, Outcome, and Study design (PICOS) framework (Higgins et al., 2019).

Inclusion criteria

Studies were eligible when they met all six substantive criteria. First, participants had to be enrolled in Kindergarten through Grade 12. Preschool-only, higher-education, and adult samples were excluded; mixed-age samples were eligible only when K–12 results could be separated. Mathematics had to be the primary

instructional subject, including numeracy, arithmetic, algebra, geometry, fractions, measurement, mathematical reasoning, or problem solving. Studies in which numerical activity was incidental to financial literacy, science, programming, or general cognitive training were excluded.

Second, the intervention had to involve a complete digital or non-digital game through which mathematics learning occurred. A complete game was operationalized as an activity with identifiable rules, a goal or end condition, player actions or decisions, challenge or progression, and feedback or consequences that formed a coherent game loop. Interventions that merely added points, badges, leaderboards, rewards, or a narrative skin to otherwise conventional tasks were treated as gamification rather than game-based learning and were excluded. Ordinary quiz platforms, technology-assisted instruction without a complete game loop, and simulations without game structure were also excluded.

Third, an eligible concurrent non-game comparator was required. Eligible comparators included business-as-usual instruction, no-treatment or wait-list conditions, active non-game practice, non-game digital instruction, and practice matched on time or content. Single-group pretest–posttest studies, historical controls, and comparisons in which all relevant arms received complete games were excluded. Comparator equivalence was treated as a moderator rather than an additional eligibility requirement.

Fourth, eligible designs were individually randomized trials, cluster-randomized trials, or quasi-experimental studies with a concurrent comparison group. Fifth, studies had to report at least one objective mathematics outcome administered comparably across conditions, such as standardized achievement, curriculum-based performance, researcher-developed mathematics tests, reasoning, problem solving, retention, or transfer. Motivation, engagement, attitudes, perceived learning, and game-only telemetry or scores were not sufficient. Sixth, the report had to provide a full account of primary research. Full journal articles, full conference papers, theses, dissertations, and sufficiently detailed research reports were eligible; reviews, protocols, editorials, abstracts, posters, and platform-development descriptions were excluded. No publication-date or language restrictions were applied. Eligibility for the systematic review was separated from eligibility for quantitative synthesis: an otherwise eligible report was retained narratively when a valid comparative effect and sampling variance could not be recovered.

Literature search and eligibility screening

The final database searches were conducted on 7 August 2026 in Scopus, Web of Science Core Collection, ERIC, and ProQuest Education Database. The strategy combined five concept blocks for complete game-based learning, mathematics, K–12 populations, objective learning outcomes, and comparative intervention designs. Terms within each block were combined with OR, and the blocks were combined with AND. Comparator equivalence and assessment proximity were not forced into the search because these characteristics required full-text interpretation and were coded after study identification. Searches were run in TITLE-ABS-KEY in Scopus, Topic in Web of Science, descriptors/free text/educational level in ERIC, and Anywhere except full text in ProQuest. No filters were used in Scopus, Web of Science, or ERIC. ProQuest was limited to Scholarly Journals or Dissertations & Theses. No date, language, subject-area, peer-review, or full-text-availability filters were applied. Complete database-specific search syntax is reported in Appendix A, consistent with PRISMA-S (Rethlefsen et al., 2021).

The searches yielded 2,633 database records, comprising 888 from Scopus, 685 from Web of Science Core Collection, 342 from ERIC, and 718 from ProQuest Education Database. Records were imported into Rayyan for screening (Ouzzani et al., 2016). Rayyan's deduplication history was independently verified. It showed that 859 duplicate records were removed before screening, leaving 1,774 records that entered and completed title–abstract screening.

Two reviewers independently assessed titles and abstracts against the predefined criteria, and disagreements were resolved through discussion until consensus was reached. Of the 1,774 screened records, 1,624 were excluded and 150 candidate reports were sought for retrieval. The primary title–abstract exclusion reasons

were wrong publication type (n = 415), wrong subject (n = 289), no eligible comparator (n = 248), wrong study design (n = 197), wrong intervention (n = 178), wrong population (n = 152), and wrong outcome (n = 145).

Full texts were sought through available library routes and, where feasible, author contact. Sixty reports remained unavailable and were recorded as reports not retrieved rather than as full-text exclusions. Ninety PDF records were available; one was an exact bibliographic/PDF duplicate and was not assessed separately. Two reviewers independently evaluated the remaining 89 unique full reports, resolving differences by discussion. Nine reports were excluded at full text: wrong intervention (n = 3), no eligible comparator (n = 4), wrong outcome (n = 1), and wrong subject (n = 1). Thus, 80 reports were retained in the systematic review. Report linkage showed that these represented 79 independent study-family units. Reports, study families, extracted effect rows, and synthesis-specific k values were tracked separately throughout the review.

Coding strategy and study-family linkage

A structured codebook was used to extract report, study, intervention, comparator, learner, outcome, and effect-size information. Before full screening, eight reports spanning clear inclusions, clear exclusions, boundary cases, a duplicate record, and a potential companion report were used to calibrate the decision rules. The codebook was then refined through documented adjudication rounds without changing eligibility according to study results. Uncertainties were resolved through reinspection of the full text and discussion between the reviewers.

Reports were linked into study families when they described the same sample, intervention implementation, and comparison design. Linkage drew on author names, country and school, recruitment and implementation dates, sample characteristics, intervention name and dosage, comparator arms, trial identifiers, and outcome timing. Exact duplicate records were retained administratively but not counted as separate reports; multiple reports from one study family were retained and used jointly for design, risk-of-bias, and outcome information.

Study-level codes included publication year, country, grade band (K–2, Grades 3–5, Grades 6–8, Grades 9–12, or mixed/unclear), design (individual RCT, cluster RCT, or quasi-experimental/unclear), and intervention format (digital, non-digital, or mixed/unclear). Mathematical outcomes were grouped as numeracy/arithmetic, algebra/fractions/advanced number, geometry/spatial, broad achievement/problem solving, or other/unclear.

Comparator conditions were coded as business-as-usual/passive, active but not fully equivalent, or equivalent active practice. Equivalence was also represented by a four-item score covering explicit matching of content, instructional time or practice dosage, technology access, and teacher attention/support. Each explicitly documented match contributed one point; absent or unclear reporting did not contribute. Assessment conditions were coded as proximal/immediate, distal/transfer/standardized, or retention/delayed. Standardization, proximity to practiced content, and time point were retained separately in the extraction file even when grouped for analysis. Multiple contrasts, shared comparator arms, outcomes, subscales, and time points were flagged as sources of dependence.

Risk-of-bias assessment

Two reviewers independently assessed risk of bias and resolved differences by consensus. Individually randomized trials were evaluated with RoB 2, cluster-randomized trials with the cluster-randomized extension of RoB 2, and non-randomized comparative studies with ROBINS-I (Sterne et al., 2016, 2019). Overall judgments followed each tool's domain-level decision rules and were not collapsed onto a common scale across tools. The assessment covered 18 individual RCT families, 17 cluster RCT families, and 44 non-randomized families. Risk of bias was not used as a full-text eligibility criterion; it informed interpretation and prespecified sensitivity analyses.

Extraction and selection of effect sizes

The unbiased standardized mean difference, Hedges' g, was the common effect metric (Hedges, 1981). Positive values were coded to favor game-based learning; lower-is-better outcomes were reverse-scored before synthesis. For raw two-group posttest data, Cohen's d was calculated using the pooled within-group standard deviation and was multiplied by the small-sample correction J. Sampling variances were calculated using standard formulas (Borenstein et al., 2009). Reported standardized effects, t or F statistics, adjusted contrasts, and model coefficients were converted to Hedges' g when the required information was available. Every derived effect retained its source table/page, calculation route, outcome, time point, contrast, and study-family identifier.

When multiple estimates represented the same contrast and outcome, the prespecified hierarchy was: an author-reported cluster-adjusted or model-based effect; a baseline-adjusted direct contrast; a gain-score standardized mean difference when the direct test targeted gain; and a raw posttest standardized mean difference for individually randomized comparisons. A validated total score was preferred to overlapping component scores in the same overall synthesis. Immediate and delayed outcomes were retained as separate, correlated effects. Pupil-level variances from cluster-assigned designs were not treated as primary unless clustering had been addressed. When cluster size was known but an adjusted variance was unavailable, design-effect scenarios using ICC values of .05, .10, and .20 were retained for sensitivity analysis only.

After de-duplication and adjudication, 286 candidate outcome rows were available across the 79 study families. The primary synthesis included 42 comparative effects from 19 families that met the primary-analysis criteria. An expanded sensitivity set added 60 effects from 26 additional families whose estimates were numerically recoverable but had limitations involving clustering, baseline adjustment, assignment, or measurement, yielding 102 effects from 45 families. Eligible families without a valid sampling variance were retained in the systematic review but were not included in the quantitative model.

Multilevel meta-analysis

Dependent effect sizes were analyzed using a three-level random-effects model rather than selecting one outcome per study or treating all effects as independent (Assink & Wibbelink, 2016; Cheung, 2014). The model included random effects for study family and effect, and parameters were estimated by restricted maximum likelihood. A working sampling variance–covariance matrix was constructed for effects within the same family using an assumed correlation of $\rho = .60$. Because the true correlations were rarely reported, sensitivity models used ρ values of 0, .30, .60, and .80. This strategy retained multiple outcomes, contrasts, shared controls, and time points while acknowledging their dependence.

To protect inference against misspecification of the working covariance structure, coefficient standard errors and confidence intervals were calculated using cluster-robust variance estimation with the CR2 small-sample adjustment and Satterthwaite degrees of freedom, clustering on study family (Hedges et al., 2010; Pustejovsky & Tipton, 2022). Models were fitted with rma.mv in metafor, and robust tests were obtained with clubSandwich (Viechtbauer, 2010). The primary model used the 42 effects from 19 families; the 102-effect, 45-family set was an expanded sensitivity analysis rather than a replacement for the primary model.

Heterogeneity was evaluated using the residual Q statistic, family- and effect-level variance components, and an approximate total I^2 calculated as the proportion of total variance attributable to the two estimated random-effect components (Higgins & Thompson, 2002). Comparator condition, assessment condition, design, digitality, grade band, mathematical outcome domain, publication year, and the equivalence score were examined in separate meta-regressions using the same covariance and random-effects structure. Univariable models were used because only 19 independent families contributed to the primary synthesis. Categorical levels represented by fewer than three families were not modeled, and a moderator model was run only when at least 10 families and at least two eligible levels remained.

Sensitivity and influence analyses

Robustness was assessed across the working-correlation values, the expanded synthesis, RCT-only and individual-RCT-only subsets, exclusion of effects with $|g| > 2$, exclusion of ROBINS-I critical-risk families from the expanded set, and leave-one-family-out analyses. A lower-concern risk-of-bias subset was prespecified but was not modeled when only four families met that criterion. Influence diagnostics were used to identify whether any family disproportionately altered the pooled estimate; they were not used to change substantive eligibility after results were known (Viechtbauer & Cheung, 2010). Cluster-variance scenarios based on ICC values of .05, .10, and .20 were kept as design-sensitivity analyses and were not substituted for the primary sampling variances.

Small-study-effect diagnostics

Small-study effects were evaluated in ways that preserved the dependent structure. First, a multilevel Egger-type meta-regression added the effect standard error as a predictor to the primary model and used CR2 inference (Egger et al., 1997; Rodgers & Pustejovsky, 2021). A parallel exploratory model used $1/\sqrt{N}$ as a precision predictor when contrast-level sample sizes were available. Second, the 42 primary effects were combined into 19 independent family summaries by generalized least squares using $\rho = .60$. These family-level summaries were used for a contour-enhanced funnel plot, PET/Egger and PEESE regressions, Begg–Mazumdar rank correlation, and trim-and-fill (Begg & Mazumdar, 1994; Duval & Tweedie, 2000; Stanley & Doucouliagos, 2014). Because these procedures had only 19 independent families and rely on different assumptions, they were treated as exploratory diagnostics of asymmetry or small-study effects rather than definitive tests of publication bias.

Software and reproducibility

Analyses were conducted in R 4.6.1 using readxl 1.5.0, metafor 5.0.1, and clubSandwich 0.7.0. The project archive contains the report–family roster, candidate-effect ledger, primary and expanded effect sets, risk-of-bias judgments, model outputs, figures, executable R scripts, report-level exclusions, family summaries, and analysis inputs.

RESULTS

Study selection and evidence structure

The four database searches yielded 2,633 records that were imported into Rayyan. Independent verification of Rayyan's deduplication history confirmed that 859 duplicate records were removed before screening. The remaining 1,774 records entered and completed title–abstract screening; 1,624 were excluded, leaving 150 candidate reports sought for retrieval.

Sixty reports could not be retrieved and were recorded as reports not retrieved. Ninety PDF records were available, of which one was an exact bibliographic and PDF duplicate that was not assessed separately. The remaining 89 unique full reports were assessed for eligibility. Nine were excluded: three for wrong intervention, four for no eligible comparator, one for wrong outcome, and one for wrong subject. Consequently, 80 reports were retained in the systematic review and were linked to 79 independent study-family units. Figure 1 presents the complete selection and counting chain.

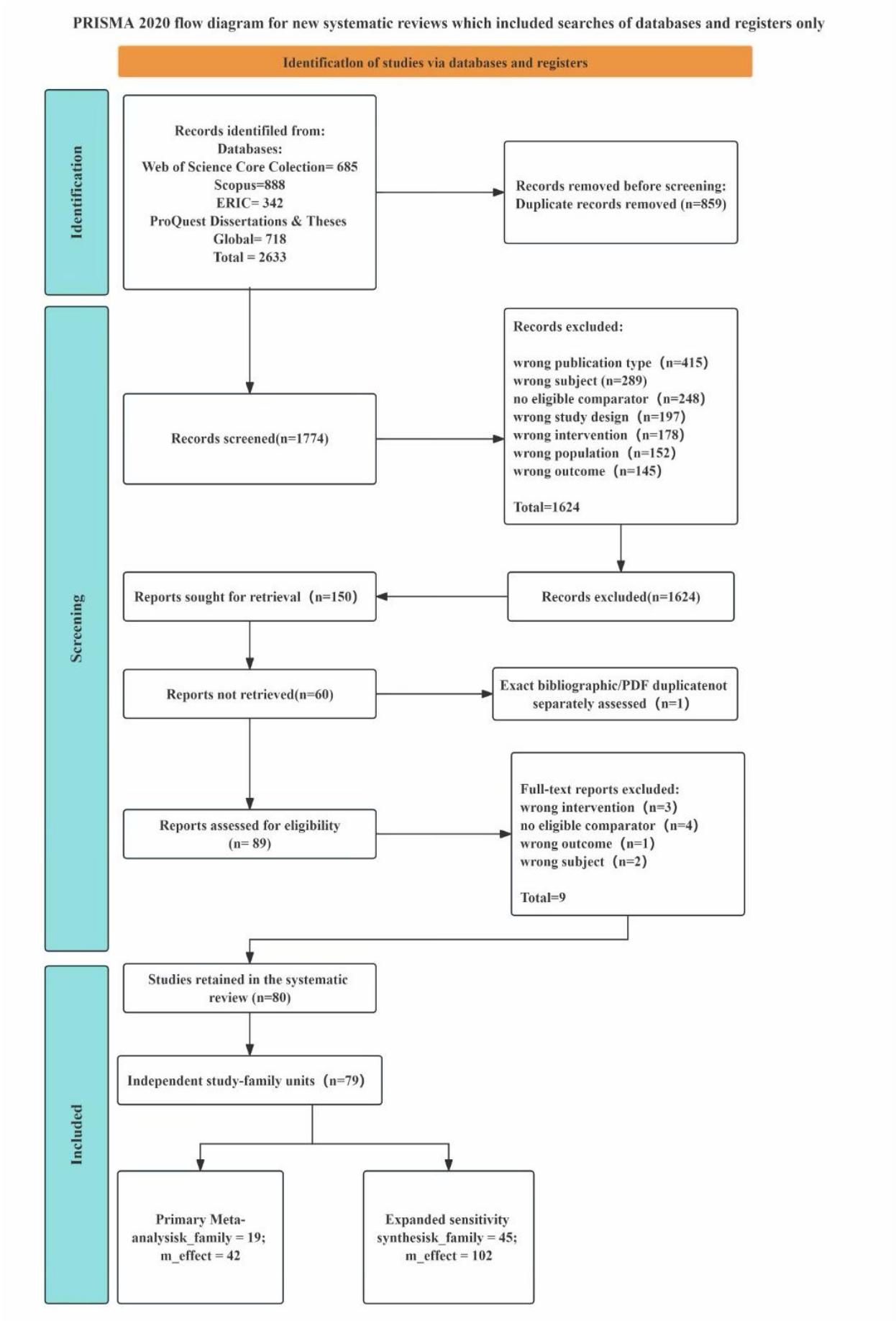

Figure 1. Study selection and evidence-counting flow. Imported records, screened records, retrieved and assessed reports, retained reports, independent study families, and synthesis-specific effect counts are shown as separate units.

Across the 79 study families, 286 de-duplicated candidate outcome rows were identified. The primary quantitative synthesis retained 42 comparative effects from 19 independent families that met the primary-analysis criteria. The expanded sensitivity set included 102 effects from 45 families. Families without a valid sampling variance remained in the systematic-review evidence base but were not included in the quantitative model.

Table 1. Selection and evidence units

Counting unit	Count	Interpretive boundary
Records imported	2,633	Database records imported into Rayyan
Records screened	1,774	After removal of 859 verified duplicate records
Candidate reports sought	150	Not final included studies
Reports not retrieved	60	Not full-text exclusions
Unique full reports assessed	89	After one exact duplicate
Reports retained	80	Systematic-review report count
Independent study families	79	Independent family units
Candidate outcome rows	286	De-duplicated effect candidates
Primary synthesis	42 effects / 19 families	Primary m_effect and k_family
Expanded sensitivity	102 effects / 45 families	Sensitivity m_effect and k_family

Note. Reports, study families, effect rows, and meta-analysis k are not interchangeable.

Study design profile and risk of bias

The 79 independent families comprised 18 individually randomized trials, 17 cluster-randomized trials, and 44 non-randomized comparative studies. Risk of bias was therefore summarized separately for RoB 2 or cluster RoB 2 and for ROBINS-I rather than collapsed onto a common scale.

Among the 35 randomized families, 29 were judged at high risk of bias and six raised some concerns; none received a low-risk overall judgment. Among the 44 non-randomized families, 10 were judged at critical risk, 32 at serious risk, and two at moderate risk; none received a low-risk overall judgment. Table 2 and Figure 2 summarize the tool-specific overall judgments. Figures 3a and 3b show the corresponding domain-level patterns for randomized and non-randomized families.

Table 2. Overall risk-of-bias judgments by assessment tool

Tool	Overall judgment	Families	Tool total
RoB 2 / cluster RoB 2	Low	0	35
RoB 2 / cluster RoB 2	Some concerns	6	35
RoB 2 / cluster RoB 2	High	29	35
ROBINS-I	Low	0	44
ROBINS-I	Moderate	2	44
ROBINS-I	Serious	32	44
ROBINS-I	Critical	10	44

Note. Judgment labels are tool-specific and should not be read as a single common low-to-high scale.

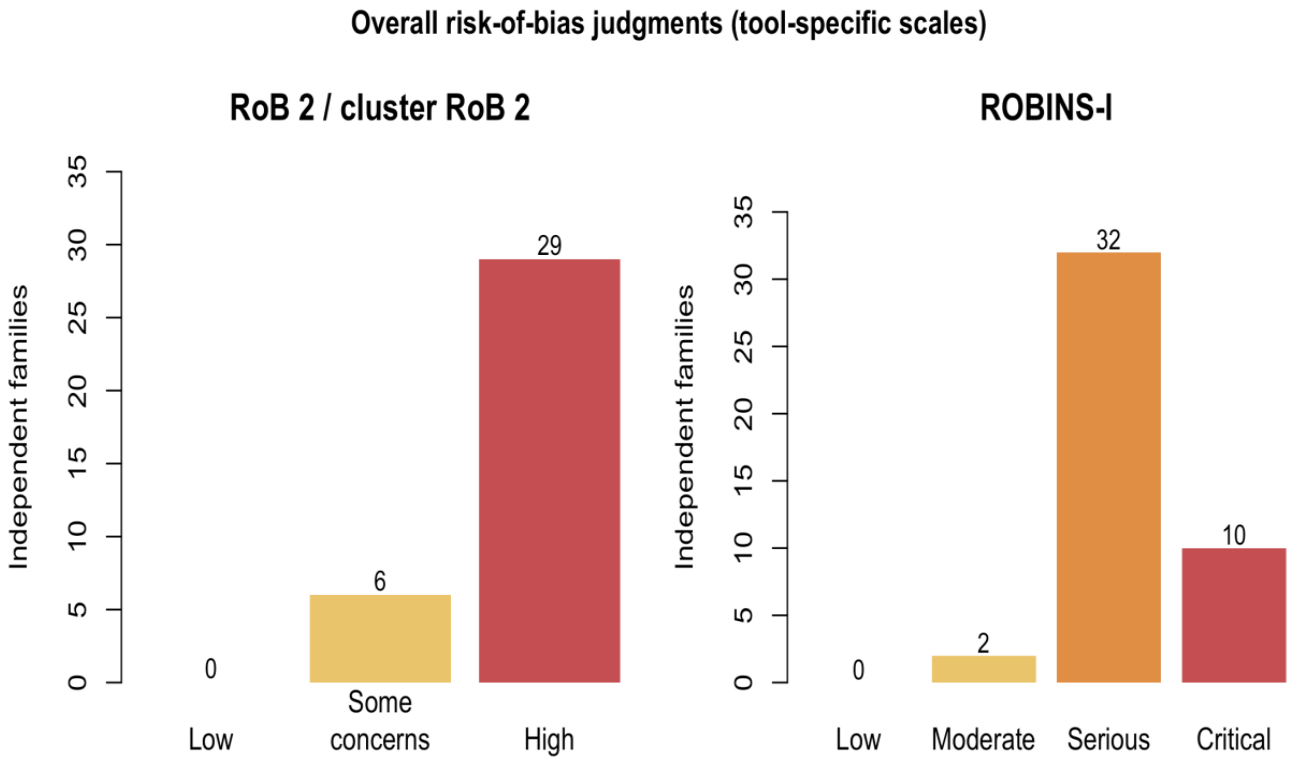

Figure 2. Overall risk-of-bias judgments shown separately for randomized and non-randomized study families.

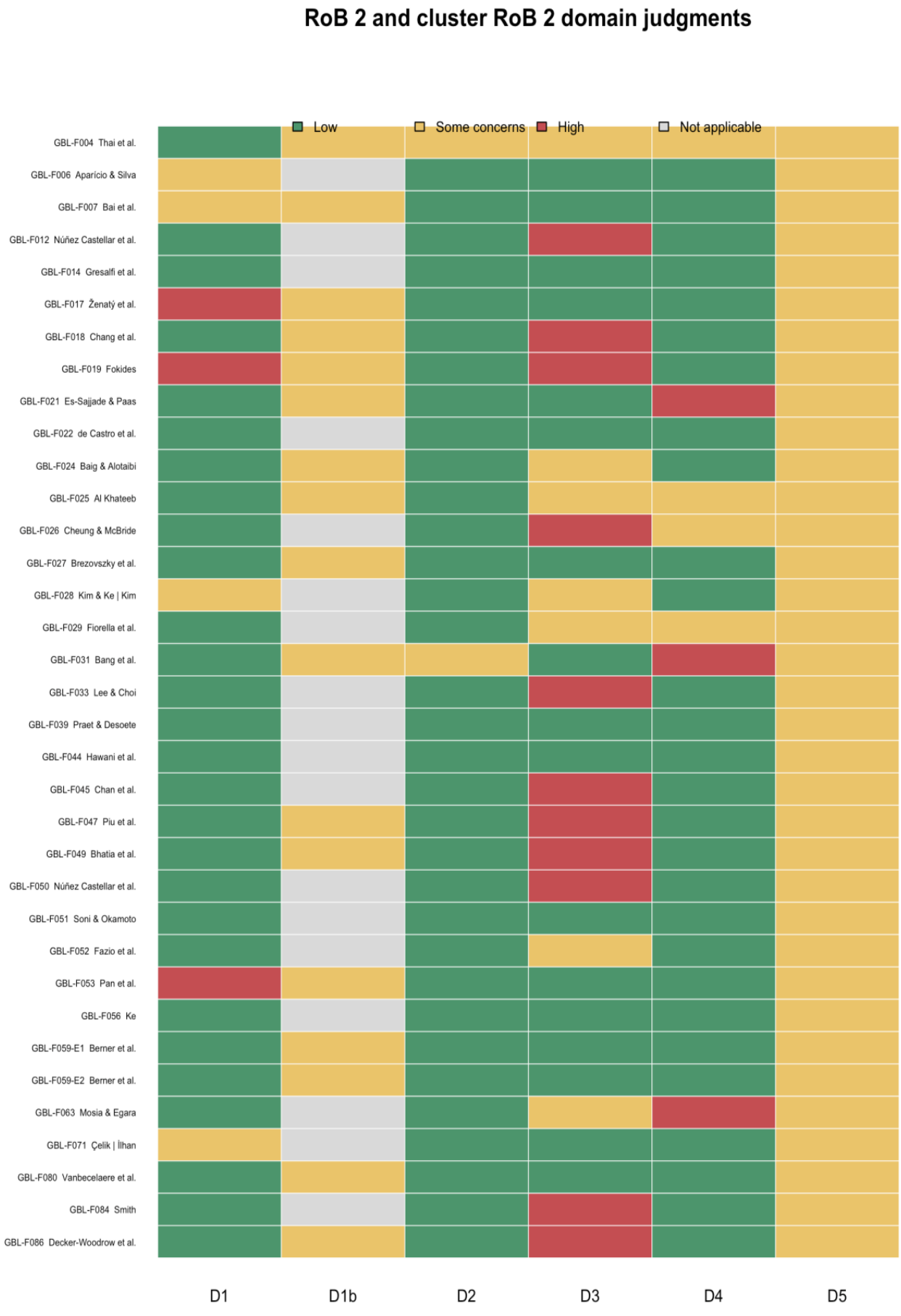

Figure 3a. Domain-level RoB 2 and cluster RoB 2 judgments for 35 randomized study families.

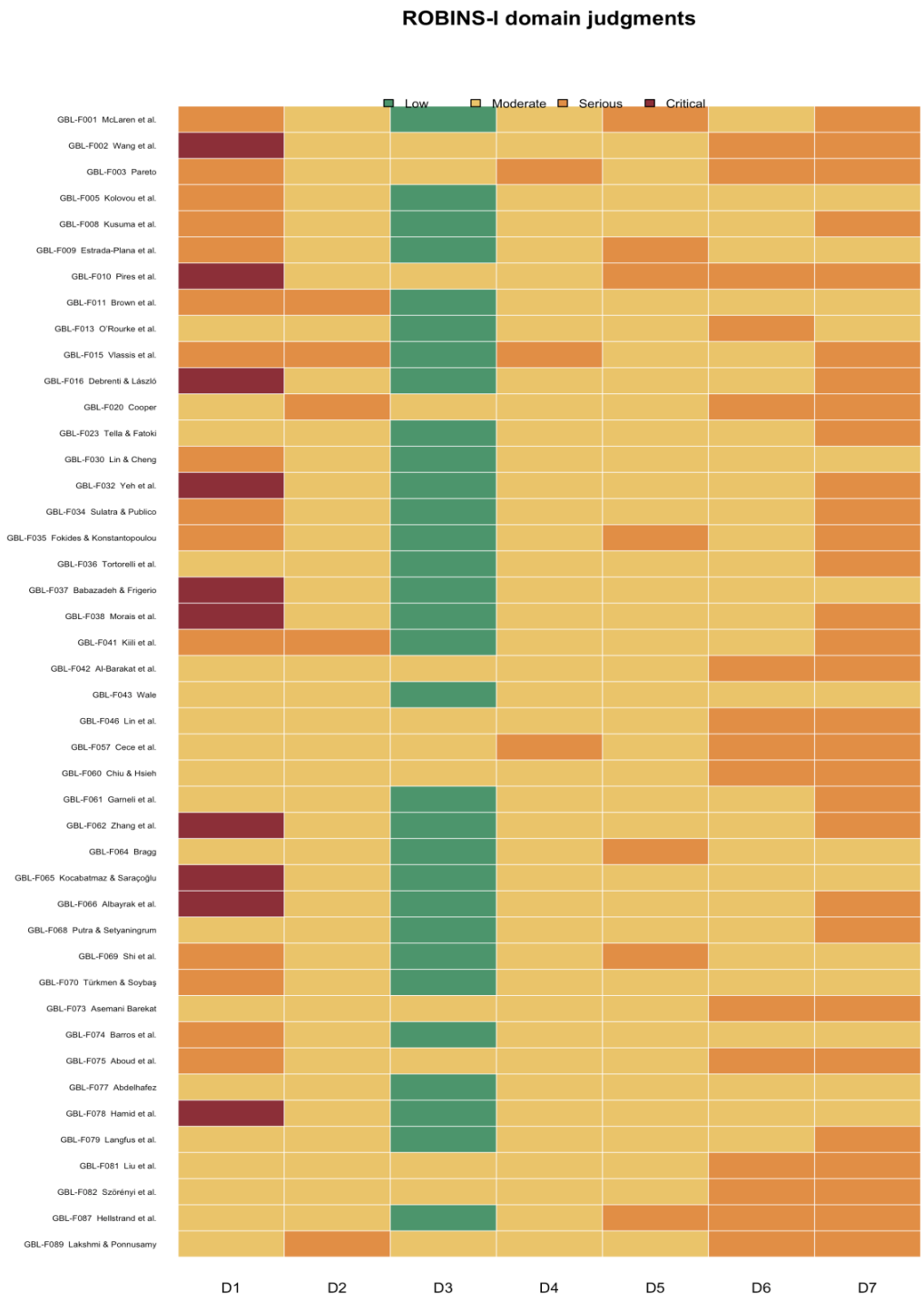

Figure 3b. Domain-level ROBINS-I judgments for 44 non-randomized study families.

Overall effect and heterogeneity (RQ1)

Table 3 presents the primary multilevel model and the working-correlation sensitivity models. The primary model used rho = .60 and included 42 dependent effects from 19 independent study families. The mean effect of game-based learning on objective mathematics outcomes was g = 0.368, SE = 0.091, 95% CI [0.176, 0.560], t(17.19) = 4.04, p < .001. Thus, the pooled estimate favored game-based learning over concurrent non-game instruction or practice.

Residual heterogeneity was substantial, Q_E(41) = 145.73, p < .001. The estimated family-level variance was tau-squared = 0.070 and the effect-level variance was tau-squared = 0.065, corresponding to an approximate total I-squared of 65.5%. Figure 4 displays the 19 independent family summaries. Its family-level random-

effects summary, g = 0.362, 95% CI [0.179, 0.545], was close to the primary multilevel estimate but served as a visualization rather than a replacement for the dependent-effect model.

Table 3. Overall multilevel models and working-correlation sensitivity

Model	g	CR2 SE	95% CI	p	Effects	Families	Approx. I2
Primary rho = 0	0.371	0.092	[.177, .564]	< .001	42	19	64.80%
Primary rho = .30	0.37	0.091	[.177, .562]	< .001	42	19	65.20%
Primary rho = .60	0.368	0.091	[.176, .560]	< .001	42	19	65.50%
Primary rho = .80	0.367	0.091	[.176, .559]	< .001	42	19	65.70%
Expanded rho = .60	0.459	0.083	[.292, .627]	< .001	102	45	83.50%

Note. The rho = .60 primary model used CR2 small-sample inference with Satterthwaite degrees of freedom. The expanded model is a sensitivity analysis.

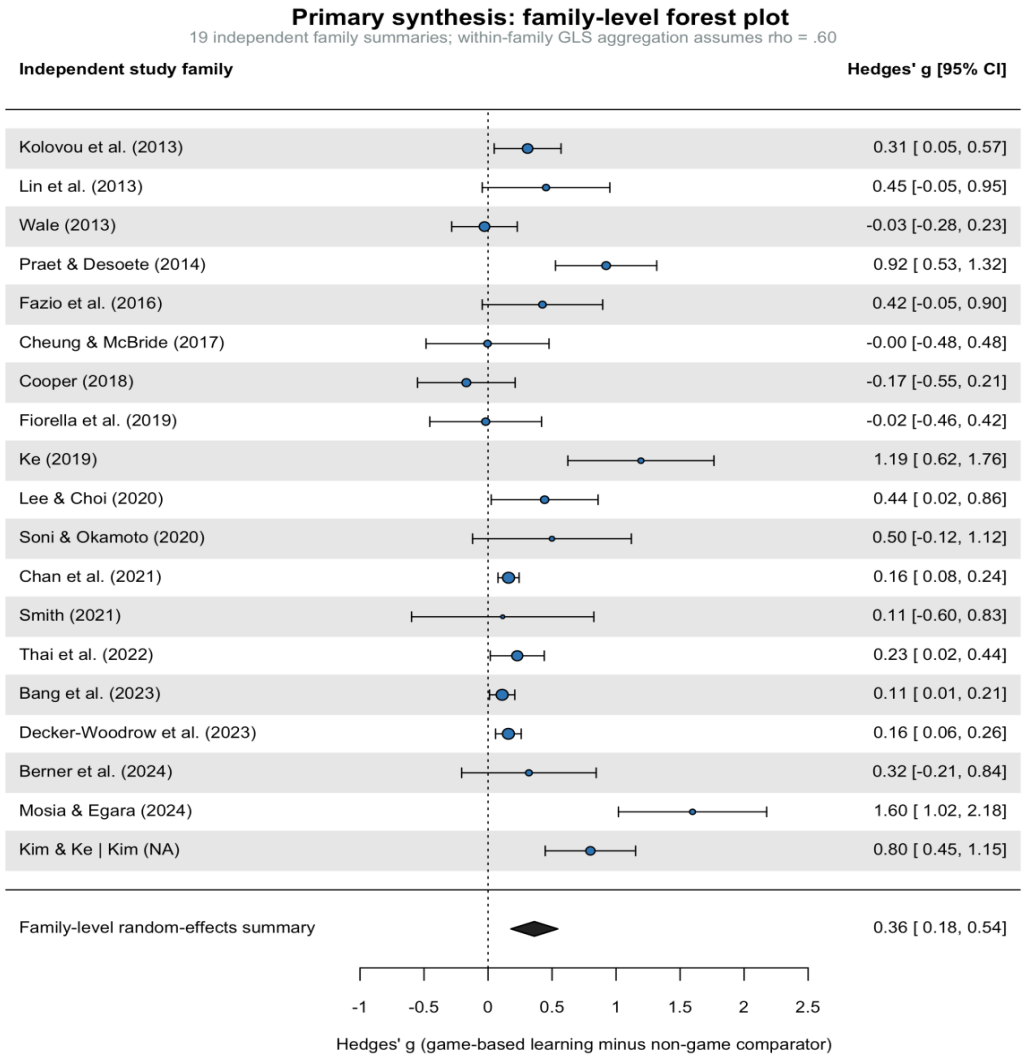

Figure 4. Family-level forest plot for the 19 independent families contributing to the primary synthesis. Within-family generalized least-squares aggregation assumes rho = .60.

Comparator and assessment conditions (RQ2-RQ3)

Table 4 presents the univariable moderator tests fitted with the same dependence and robust-inference structure as the primary model. Comparator condition did not significantly moderate the effects, Q_M(2) = 1.402, p = .496. With active but not fully equivalent practice as the reference category, the coefficient for business-as-usual or passive comparators was b = -0.138, 95% CI [-0.900, 0.623], p = .683, and the coefficient for equivalent active practice was b = -0.252, 95% CI [-1.008, 0.504], p = .460. The continuous equivalence score likewise was not associated with effect size, b = -0.008 per point, 95% CI [-0.124, 0.107], p = .856.

For assessment condition, the modelable contrast compared proximal or immediate outcomes with distal, transfer, or standardized outcomes. The omnibus test did not reach statistical significance, Q_M(1) = 2.314, p = .128. The proximal or immediate coefficient relative to the distal reference category was positive but imprecise, b = 0.256, 95% CI [-0.088, 0.600], p = .120. These analyses therefore did not provide clear evidence that either comparator equivalence or assessment proximity explained the observed heterogeneity.

Other moderators (RQ4)

None of the other prespecified moderator sets produced a statistically significant omnibus test. The results were Q_M(2) = 4.273, p = .118 for study design; Q_M(1) = 0.477, p = .490 for intervention digitality; Q_M(2) = 3.140, p = .208 for grade band; Q_M(3) = 1.227, p = .747 for mathematics outcome domain; and Q_M(1) = 0.038, p = .845 for publication year. Grade band was estimated from 35 effects in 16 families, outcome domain from 41 effects in 18 families, and publication year from 41 effects in 18 families because of missing or non-modelable categories. Given the small and uneven cells, these tests were inconclusive and should not be interpreted as evidence of equivalence across categories.

Table 4. Univariable moderator omnibus tests

Moderator	QM (df)	p	Effects	Families	Result
Comparator condition	1.402 (2)	0.496	42	19	Not statistically significant
Assessment condition	2.314 (1)	0.128	42	19	Not statistically significant
Study design	4.273 (2)	0.118	42	19	Not statistically significant
Digitality	0.477 (1)	0.49	42	19	Not statistically significant
Grade band	3.140 (2)	0.208	35	16	Not statistically significant
Mathematics outcome domain	1.227 (3)	0.747	41	18	Not statistically significant
Publication year	0.038 (1)	0.845	41	18	Not statistically significant
Equivalence score	0.033 (1)	0.855	42	19	Not statistically significant

Note. Models were univariable. Family counts can differ because some moderator levels were missing or not modelable.

Sensitivity and influence analyses (RQ5)

The primary estimate was stable across the assumed within-family correlations: pooled effects ranged from g = 0.367 to 0.371 for rho values from 0 to .80, and all confidence intervals excluded zero. The expanded 102-effect, 45-family model produced a larger estimate, g = 0.459, SE = 0.083, 95% CI [0.292, 0.627], p < .001, with greater approximate heterogeneity (I-squared = 83.5%). Because the expanded set included effects with

additional design, clustering, baseline-adjustment, or measurement limitations, it was treated as a sensitivity analysis rather than a replacement for the primary model.

Restriction to randomized designs yielded g = 0.432, SE = 0.108, 95% CI [0.200, 0.665], p = .001, across 38 effects from 15 families. Restriction to individually randomized trials yielded g = 0.541, SE = 0.147, 95% CI [0.213, 0.869], p = .004, across 32 effects from 11 families. Excluding effects with absolute g greater than 2 did not change the primary result, and excluding ROBINS-I critical-risk families did not change the expanded result. The prespecified lower-concern risk-of-bias subset was not modeled because only four families met that threshold.

In the 19 leave-one-family-out models, pooled estimates ranged from g = 0.312 to 0.395. Every estimate remained positive and every confidence interval excluded zero. Figure 5 shows that no single family reversed the direction or statistical conclusion of the primary synthesis.

Table 5. Design and risk-of-bias sensitivity analyses

Analysis	g	CR2 SE	95% CI	p	Effects	Families
Primary: RCT designs only	0.432	0.108	[.200, .665]	0.001	38	15
Primary: individual RCT only	0.541	0.147	[.213, .869]	0.004	32	11
Primary: absolute g <= 2	0.368	0.091	[.176, .560]	< .001	42	19
Expanded: exclude ROBINS-I critical	0.459	0.083	[.292, .627]	< .001	102	45
Primary: lower-concern RoB	Not modeled	-	-	-	-	4 eligible families

Note. The lower-concern subset was too small for the prespecified model.

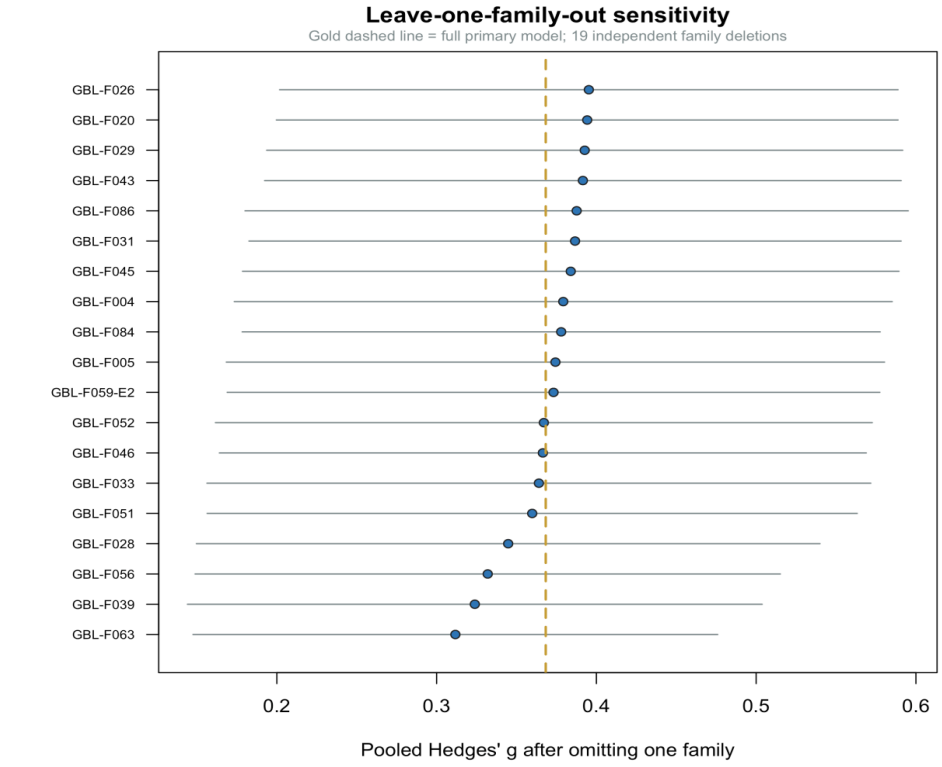

Figure 5. Leave-one-family-out sensitivity analysis. The dashed line marks the full primary estimate; points and intervals show the estimate after each family was omitted.

Small-study-effect diagnostics (RQ5)

The multilevel Egger-type model indicated a positive association between effect size and standard error, b = 1.775, SE = 0.740, 95% CI [0.098, 3.453], t(8.89) = 2.40, p = .040. In contrast, the parallel sample-size precision coefficient was not statistically significant, b = 2.217, 95% CI [-1.009, 5.443], p = .156. At the independent family level, the PET/Egger slope was b = 1.874, p = .041, and the PEESE variance slope was b = 5.015, p = .046. The rank-correlation test was close to but did not reach conventional significance, Kendall's tau = .322, p = .058.

The unadjusted family-level random-effects estimate was g = 0.362, 95% CI [0.179, 0.545]. Trim-and-fill imputed three family summaries and produced an adjusted estimate of g = 0.225, 95% CI [-0.012, 0.461], p = .062. Figure 6 displays the contour-enhanced funnel plot. Taken together, the SE-based regressions suggested possible asymmetry or small-study effects, whereas the sample-size and rank-correlation diagnostics were less conclusive. Given the 19-family denominator and the different assumptions of these procedures, the pattern was treated as an exploratory signal rather than definitive evidence of publication bias.

Table 6. Exploratory small-study-effect diagnostics

Diagnostic	Statistic / estimate	95% CI	p	Independent unit
Multilevel Egger SE slope	b = 1.775	[0.098, 3.453]	0.04	42 effects / 19 families
Multilevel 1/sqrt(N) slope	b = 2.217	[-1.009, 5.443]	0.156	42 effects / 19 families
Family PET/Egger slope	b = 1.874	[0.073, 3.674]	0.041	19 family summaries
Family PEESE variance slope	b = 5.015	[0.087, 9.943]	0.046	19 family summaries
Rank correlation	tau = .322	-	0.058	19 family summaries
Trim-and-fill adjusted g	.225; k0 = 3	[-.012, .461]	0.062	22 observed/imputed summaries

Note. All diagnostics are exploratory. The trim-and-fill estimate does not replace the primary multilevel model.

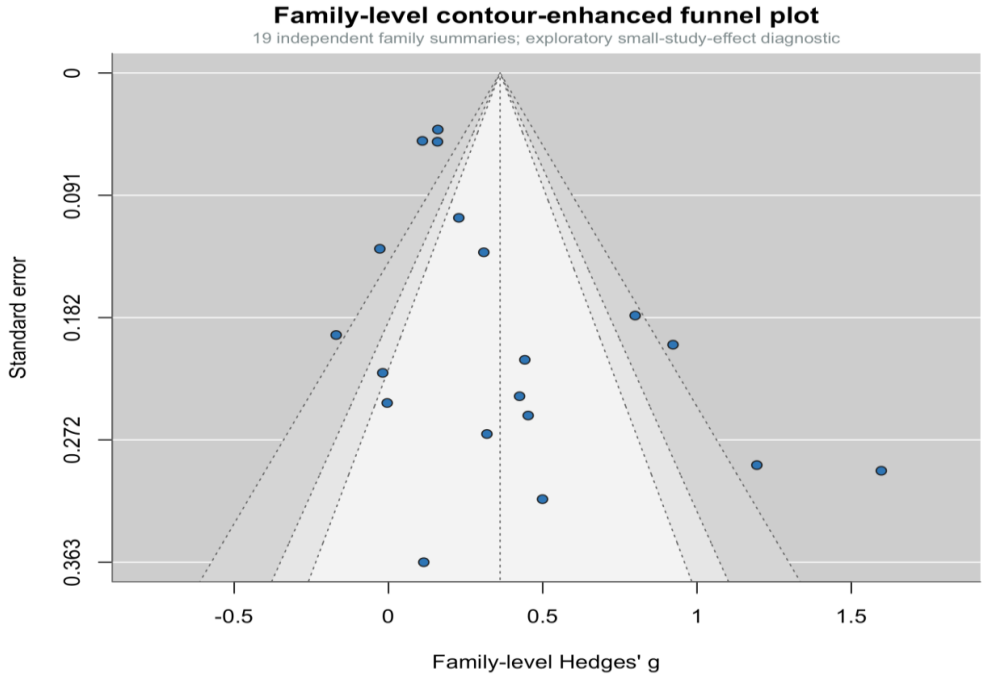

Figure 6. Contour-enhanced funnel plot of 19 independent family summaries. The plot is an exploratory diagnostic of asymmetry and small-study effects.

DISCUSSION

Principal findings and overall effect

This systematic review identified 80 eligible reports linked to 79 independent study families, whereas the primary quantitative synthesis comprised 42 effects from 19 independent families that met the primary-analysis criteria. Across those effects, complete digital or non-digital game-based learning (GBL) produced a positive average effect on objective K–12 mathematics outcomes relative to concurrent non-game instruction or practice, g = 0.368, 95% CI [0.176, 0.560]. The estimate represents a modest-to-moderate standardized advantage, although its educational importance also depends on the counterfactual, outcome measure, instructional duration, implementation demands, and costs (Kraft, 2020). For RQ1, complete mathematics GBL was more effective on average than the non-game conditions represented in the primary synthesis.

The direction and approximate magnitude of this finding are consistent with earlier syntheses of digital games and mathematics learning. Clark et al. (2016), for example, reported an average game-versus-non-game effect of g = 0.33 across learning domains, and Byun and Joung (2018) reported a positive average effect for digital K–12 mathematics games. Mathematics-focused and serious-game syntheses have likewise generally favored game conditions, although their estimates and moderator findings have varied (Tokac et al., 2019; Wouters et al., 2013). These comparisons should not be treated as replications of an identical effect. Earlier reviews differed in whether they included only digital games, pre-K or postsecondary learners, gamified tasks, simulations, affective outcomes, or studies without the same comparator requirements. The present estimate is therefore best understood as convergent evidence from a construct-strict K–12 mathematics synthesis, not as a pooled update of interchangeable reviews.

The positive mean is theoretically compatible with several ways in which a complete game can organize mathematical activity. Goal–action–feedback cycles can create repeated opportunities to make decisions, observe consequences, and revise strategies (Garris et al., 2002). Games can also coordinate cognitive, motivational, affective, and social supports (Plass et al., 2015), particularly when the mathematics is intrinsically integrated with the actions required for progress (Habgood & Ainsworth, 2011). However, these mechanisms were not tested as mediators in the present meta-analysis. The findings do not show that engagement, narrative, competition, technology, or feedback individually caused the achievement advantage, and they do not imply that the label “game” is itself an active ingredient.

The pooled estimate also concealed substantial residual heterogeneity, $Q_E(41) = 145.73$, $p < .001$, with an approximate total I^2 of 65.5%. This variation is unsurprising because the included interventions differed in game design, mathematics content, learner age, instructional setting, comparator strength, assessment alignment, and study design. Explaining that variation was a central purpose of the review. Yet the primary moderator analyses were based on only 19 independent families, and several categories were represented by still fewer families. Given these constraints, the moderator estimates show what the available evidence established, but they do not provide a definitive account of when mathematics games work best.

Comparator and study-design conditions

Comparator condition did not significantly moderate effect sizes, $Q_M(2) = 1.402$, $p = .496$, and the continuous comparator-equivalence score was also unrelated to the observed effects, $b = -0.008$ per point, 95% CI [−0.124, 0.107], $p = .856$. The coefficients for business-as-usual/passive and equivalent active-practice comparisons were imprecise, with confidence intervals spanning both potentially meaningful advantages and disadvantages. These results therefore do not demonstrate that comparator strength is unimportant. Meta-analytic moderator tests with few studies and uneven subgroup representation can have limited power, and confidence intervals are essential for interpreting what remains plausible (Valentine et al., 2010).

The pooled effect pertains to the concurrent non-game comparators represented in the primary model. It does not estimate the incremental effect of organizing otherwise identical mathematics practice as a game. A comparison with business-as-usual instruction can capture differences in content, dosage, technology, feedback, teacher attention, or novelty in addition to game structure. Equivalent active practice holds these

features more nearly constant and offers a stricter value-added test of the game format. Prior methodological work shows that comparison quality and other design features affect how educational effects should be interpreted (Cheung & Slavin, 2016; Clark et al., 2016; Kraft, 2020). Equivalent-practice evidence in the present review was sparse and imprecise. The results therefore support an average advantage over the available non-game alternatives while leaving the added value over fully matched practice unresolved.

Study design also failed to produce a statistically significant omnibus test, $QM(2) = 4.273$, $p = .118$, and therefore did not support the expectation that quasi-experimental designs would yield larger estimates. Restriction to randomized designs retained a positive effect, $g = 0.432$, 95% CI [0.200, 0.665], and restriction to individually randomized trials yielded $g = 0.541$, 95% CI [0.213, 0.869]. These sensitivity estimates strengthen confidence that the positive direction was not produced solely by the non-randomized studies in the primary set. They do not show that randomized games are more effective than quasi-experimental games: the restricted models contain different, overlapping subsets of interventions, and the design moderator itself was not significant.

Randomization also did not remove all credibility concerns. None of the 35 randomized families received a low-risk overall judgment; 29 were rated high risk and six raised some concerns. Among the 44 non-randomized families, 42 were at serious or critical risk under ROBINS-I. These patterns mean that a positive design-restricted estimate should not be interpreted as a bias-free causal effect. Future evaluations need both stronger allocation procedures and better protection against missing outcome data, outcome-selection bias, deviations from intended intervention, and confounding.

Assessment conditions

The assessment-condition analysis did not establish a difference between proximal/immediate and distal, transfer, or standardized outcomes, $QM(1) = 2.314$, $p = .128$. Nevertheless, the proximal/immediate coefficient was positive, $b = 0.256$, 95% CI [−0.088, 0.600], $p = .120$. The direction is compatible with the a priori expectation that measures closely aligned with the taught content would be more sensitive to intervention effects, but the confidence interval includes both little difference and a substantively larger proximal effect. It would therefore be incorrect to conclude either that treatment–outcome alignment explains the pooled result or that GBL transfers equally well to more independent assessments.

Assessment proximity determines what kind of learning claim an effect can support. A proximal test can validly detect mastery of the specific content practiced in the game, yet a large effect on such a test does not necessarily generalize to unfamiliar problems or standardized achievement. Distal, transfer, and standardized measures provide a more demanding test, although they may be less sensitive to the exact knowledge taught. Methodological reviews have documented larger effects for researcher-developed than independent measures, underscoring the importance of separating instructional impact from treatment–measure alignment (Cheung & Slavin, 2016; Kraft, 2020). In the present review, the modelable categories combined several assessment properties, and delayed outcomes were too sparse for a separate robust moderator claim. Future studies should therefore administer both a well-aligned proximal measure and an independently developed distal or transfer measure, with delayed follow-up when retention is a stated goal.

Other intervention, learner, and outcome moderators

The remaining prespecified moderators—intervention digitality, grade band, mathematics outcome domain, and publication year—did not reach statistical significance. Digitality yielded $QM(1) = 0.477$, $p = .490$, suggesting no established advantage for either digital or non-digital games in the available primary data. This finding supports keeping both media within the construct of complete GBL, but it does not prove that they are equally effective. Digital games can automate adaptation, feedback, and representation, whereas non-digital games can make rules and quantities tangible and support face-to-face explanation. Whether either medium improves mathematics depends on the design and enactment of the activity, not on medium alone.

A more useful research question is therefore which combinations of mechanics, representations, feedback, scaffolding, and teacher facilitation support particular mathematical goals. Reviews of mathematics games

document substantial variation in pedagogical approach and knowledge type (Kaçmaz & Dubé, 2022). Ke (2016) further emphasizes the purposeful integration of learning actions with game actions, iterative reflection, and in-game support, while Wouters and van Oostendorp (2013) show that instructional support can improve learning in complex GBL environments. These sources offer plausible explanations for between-study variation, but the current meta-analysis could not model these features consistently because primary reports rarely described them in sufficient detail.

Grade band did not significantly moderate effects, $QM(2) = 3.140$, $p = .208$, but this model contained only 35 effects from 16 families. The absence of a detected difference cannot support a developmental claim that the same game structures work equally well from Kindergarten through secondary school. Similarly, mathematics outcome domain was not a significant moderator, $QM(3) = 1.227$, $p = .747$, and publication year showed no linear association, $QM(1) = 0.038$, $p = .845$. These results suggest that a positive average is not confined to one obvious medium, age band, domain, or publication era; however, the coarse categories and small cells leave open interactions among learner knowledge, game design, mathematical demand, and classroom implementation.

Robustness, risk of bias, and small-study effects

Several analyses supported the stability of the positive primary estimate. Varying the assumed within-family correlation from 0 to .80 changed the pooled effect only from $g = 0.371$ to 0.367, and all corresponding confidence intervals excluded zero. The expanded sensitivity set of 102 effects from 45 families also favored GBL, $g = 0.459$, although heterogeneity increased to 83.5% and the set included additional design, clustering, baseline-adjustment, or measurement limitations. Leave-one-family-out estimates ranged from $g = 0.312$ to 0.395, and no single family changed the direction or statistical conclusion. Together, these checks indicate that the primary mean was not an artifact of the chosen working correlation or one unusually influential family.

Analytic robustness did not remove concerns about study quality. The tool-specific judgments showed pervasive concerns across randomized and non-randomized families, and only four families met the prespecified lower-concern threshold—too few for the planned restricted model. The expanded model remained positive after excluding ROBINS-I critical-risk families, but many serious-risk studies remained. This limitation is especially important because second-order evidence on mathematics game research has linked review conclusions to the quality of the underlying evidence base (Yao, 2026). The pooled estimate should therefore be interpreted cautiously because the evidence base remains at substantial risk of bias.

The small-study-effect diagnostics added a further caution. Standard-error-based multilevel and family-level regressions were statistically significant, and the PEESE slope was also significant. In contrast, the sample-size precision coefficient and rank-correlation test were not significant at conventional levels. Trim-and-fill imputed three family summaries and reduced the family-level estimate from $g = 0.362$ to $g = 0.225$, 95% CI [−0.012, 0.461], $p = .062$. These procedures rely on different assumptions and were applied to only 19 independent family summaries. The mixed pattern is consistent with possible funnel asymmetry or small-study effects, but it cannot distinguish publication bias from heterogeneity, design differences, measurement alignment, or chance. The adjusted trim-and-fill estimate is therefore a sensitivity signal, not a replacement for the primary multilevel model.

Strengths and limitations

The review defined the intervention as a complete game and separated GBL from gamification and ordinary technology-assisted instruction. It included digital and non-digital games, required a concurrent non-game comparator, linked multiple reports to independent study families, and kept reports, families, effect sizes, and synthesis-specific k values distinct. The analysis retained dependent outcomes in a multilevel model with CR2 small-sample inference, explicitly coded comparator equivalence and assessment condition, used design-appropriate risk-of-bias tools, and examined working correlations, synthesis sets, design restrictions, influential families, and multiple small-study diagnostics.

Several limitations qualify the findings. First, 60 of the 150 candidate reports sought for retrieval were unavailable and were recorded as reports not retrieved. Their eligibility and results are unknown, and no assumption can be made that the retrieved evidence is representative of them. The review therefore reflects the available full reports rather than every potentially eligible report identified at title–abstract screening.

Second, the quantitative evidence base was much smaller than the systematic-review evidence base. Although 80 reports and 79 independent families were retained, only 19 families contributed effects with sampling variances meeting the primary-analysis criteria; 45 families contributed to the broader sensitivity set. Excluding non-computable results avoids assigning unsupported precision, but it can create availability bias if complete statistical reporting is related to study design or results. The primary estimate should not be presented as if all 79 families contributed equal quantitative information.

Third, coding depended on the detail and terminology of primary reports. Descriptions of comparator content, instructional time, practice dosage, technology access, teacher support, game mechanics, adaptation, implementation fidelity, and assessment alignment were often incomplete. Unclear matching features were not assumed to be equivalent, but this conservative rule cannot recover information that was never reported. Broad moderator categories may consequently combine substantively different conditions and attenuate real interactions.

Fourth, the moderator analyses were univariable and involved few independent families, uneven category sizes, and residual heterogeneity. Non-significant omnibus tests therefore provide limited evidence against moderation, especially for grade band, outcome domain, and equivalent active practice. The study-level coding also could not test individual-level interactions involving prior knowledge, disability, language, socioeconomic context, or engagement. Nor can aggregate effect sizes establish whether motivation, feedback, strategic reflection, time on task, or intrinsic integration mediated mathematics learning.

Finally, risk of bias was high across much of the primary literature, and the small-study diagnostics were exploratory. Cluster trials often required assumptions or corrections because intraclass correlations and cluster-adjusted statistics were incompletely reported. The primary model was stable across the prespecified correlation values, but this does not validate every effect-size input or covariance assumption. These limitations collectively argue for cautious causal language and for replication with better reported, independently measured, and rigorously randomized comparisons.

Implications for research and practice

For practice, complete mathematics GBL can be considered one instructional option, but the findings do not support its universal replacement of non-game teaching. A game is most appropriate when progress requires the target mathematics, feedback is interpretable, challenge is calibrated to learners, and gameplay is integrated with explanation, reflection, or teacher facilitation. Intrinsic integration and purposeful learning–play design are more actionable principles than adding game features without a mathematical function (Habgood & Ainsworth, 2011; Ke, 2016). Instructional support may be particularly important when a game presents a complex environment in which learners could otherwise attend to irrelevant information or fail to connect actions with mathematical ideas (Wouters & van Oostendorp, 2013).

The non-significant digitality moderator permits flexibility but not a claim of equal effectiveness. Educators can select digital or non-digital games according to the mathematical goal, learner needs, classroom interaction, accessibility, teacher expertise, and available resources. Likewise, the positive pooled effect does not justify replacing a well-designed non-game practice activity solely because a game is more engaging. Where the practical question is whether the game format adds value, the relevant local comparison is an activity that matches content, time, practice, feedback, technology, and teacher attention as closely as possible.

For research, the priority is a stronger counterfactual. Future trials should use individual or cluster randomization where feasible and compare complete games with deliberately matched non-game practice. Factorial or value-added designs can isolate particular mechanics, feedback systems, narratives, collaborative structures, or scaffolds. Studies should report allocation and attrition, cluster sizes and intraclass correlations,

baseline and posttest statistics, adjusted effect estimates, all prespecified outcomes, and the exact content and dosage received by every condition.

Measurement should be designed as part of the causal question. At minimum, evaluations should pair a proximal measure that is sensitive to the taught mathematics with a distal or independently developed measure that tests generalization; delayed follow-up should be added when retention is claimed. Preregistration, protocol-publication, complete outcome reporting, and sharable de-identified data would reduce selective flexibility and improve effect-size recovery. Intervention and comparator descriptions should cover materials, procedures, providers, setting, dosage, tailoring, modifications, and fidelity, consistent with the logic of the TIDieR reporting framework (Hoffmann et al., 2014). Such studies would make it possible to move beyond the broad question of whether games work and identify when a complete game contributes learning value beyond an equally intensive alternative.

CONCLUSION

Complete digital and non-digital GBL interventions were associated with better objective K–12 mathematics outcomes than concurrent non-game instruction or practice. In the primary multilevel model, 42 effects from 19 independent study families yielded $g = 0.368$, 95% CI [0.176, 0.560]. The positive direction persisted across working-correlation assumptions, randomized-design restrictions, the expanded synthesis set, and every leave-one-family-out analysis.

The result is encouraging but narrower than a general claim that games improve mathematics beyond equivalent practice. Comparator category, comparator-equivalence score, assessment condition, design, digitality, grade band, mathematics domain, and publication year were not established as moderators, but the small and uneven evidence base prevents those null tests from demonstrating equivalence. Most study families had serious credibility concerns, 60 candidate reports were not retrieved, only 19 of 79 retained families contributed to the primary quantitative model, and exploratory diagnostics suggested possible small-study effects.

Overall, complete mathematics GBL was associated with higher objective achievement than the non-game alternatives represented in the available literature. Whether the game format adds value over fully matched practice remains unresolved. Future trials should use randomized designs, match comparison conditions, include both proximal and transfer outcomes, and report game mechanisms and instructional supports in sufficient detail.

DECLARATIONS

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Competing interests

The authors declare that they have no competing interests.

Ethics approval

Not applicable. This systematic review and meta-analysis synthesized data from previously conducted studies and did not involve the collection of new data from human participants or animals; institutional ethics approval was therefore not required.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Welfare of animals

Not applicable.

REFERENCES

1. Anggoro, B. S., Dewantara, A. H., Suherman, S., Muhammad, R. R., & Saraswati, S. (2025). Effect of game-based learning on students' mathematics high order thinking skills: A meta-analysis. Revista de Psicodidáctica, 30(1), 500158. https://doi.org/10.1016/j.psicoe.2024.500158
2. Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. The Quantitative Methods for Psychology, 12(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154
3. 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
4. Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. Biometrics, 50(4), 1088–1101. https://doi.org/10.2307/2533446
5. Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. Wiley. https://doi.org/10.1002/9780470743386
6. 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
7. Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. Educational Researcher, 45(5), 283–292. https://doi.org/10.3102/0013189X16656615
8. Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. Psychological Methods, 19(2), 211–229. https://doi.org/10.1037/a0032968
9. Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. Review of Educational Research, 86(1), 79–122. https://doi.org/10.3102/0034654315582065
10. Conmy, T. (2023). We are still playing: A meta-analysis of game-based learning in mathematics education [Doctoral dissertation]. ProQuest Dissertations & Theses Global. (ERIC No. ED645006)
11. Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining gamification. In Proceedings of the 15th International Academic MindTrek Conference (pp. 9–15). ACM. https://doi.org/10.1145/2181037.2181040
12. 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
13. Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. BMJ, 315, 629–634. https://doi.org/10.1136/bmj.315.7109.629
14. Ersen, Z. B., & Ergül, E. (2022). Trends of game-based learning in mathematics education: A systematic review. International Journal of Contemporary Educational Research, 9(3), 603–623. https://doi.org/10.33200/ijcer.1109501
15. Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. Simulation & Gaming, 33(4), 441–467. https://doi.org/10.1177/1046878102238607
16. Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. Journal of the Learning Sciences, 20(2), 169–206. https://doi.org/10.1080/10508406.2010.508029
17. Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6(2), 107–128. https://doi.org/10.3102/10769986006002107
18. Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods, 1(1), 39–65. https://doi.org/10.1002/jrsm.5

19. Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21(11), 1539–1558. https://doi.org/10.1002/sim.1186
20. Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). Cochrane handbook for systematic reviews of interventions (2nd ed.). Wiley. https://doi.org/10.1002/9781119536604
21. Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D. G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A.-W., & Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. BMJ, 348, g1687. https://doi.org/10.1136/bmj.g1687
22. Hui, H. B., & Mahmud, M. S. (2023). Influence of game-based learning in mathematics education on the students' cognitive and affective domain: A systematic review. Frontiers in Psychology, 14, 1105806. https://doi.org/10.3389/fpsyg.2023.1105806
23. Kaçmaz, G., & Dubé, A. K. (2022). Examining pedagogical approaches and types of mathematics knowledge in educational games: A meta-analysis and critical review. Educational Research Review, 35, 100428. https://doi.org/10.1016/j.edurev.2021.100428
24. Ke, F. (2016). Designing and integrating purposeful learning in game play: A systematic review. Educational Technology Research and Development, 64(2), 219–244. https://doi.org/10.1007/s11423-015-9418-1
25. Kraft, M. A. (2020). Interpreting effect sizes of education interventions. Educational Researcher, 49(4), 241–253. https://doi.org/10.3102/0013189X20912798
26. Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. Educational Psychology Review, 22(3), 215–243. https://doi.org/10.1007/s10648-010-9125-8
27. National Council of Teachers of Mathematics. (2014). Principles to actions: Ensuring mathematical success for all. Author.
28. Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. Systematic Reviews, 5, 210. https://doi.org/10.1186/s13643-016-0384-4
29. Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
30. 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
31. Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. Prevention Science, 23(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3
32. Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., et al. (2021). PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Systematic Reviews, 10, 39. https://doi.org/10.1186/s13643-020-01542-z
33. Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. Psychological Methods, 26(2), 141–160. https://doi.org/10.1037/met0000300
34. Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. Educational Psychology Review, 32, 77–112. https://doi.org/10.1007/s10648-019-09498-w
35. Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. Research Synthesis Methods, 5(1), 60–78. https://doi.org/10.1002/jrsm.1095
36. Sterne, J. A. C., Hernán, M. A., Reeves, B. C., et al. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ, 355, i4919. https://doi.org/10.1136/bmj.i4919
37. Sterne, J. A. C., Savović, J., Page, M. J., et al. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ, 366, l4898. https://doi.org/10.1136/bmj.l4898
38. 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
39. Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. Journal of Educational and Behavioral Statistics, 35(2), 215–247. https://doi.org/10.3102/1076998609346961

40. Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1–48. https://doi.org/10.18637/jss.v036.i03
41. 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
42. Wouters, P., & van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. Computers & Education, 60(1), 412–425. https://doi.org/10.1016/j.compedu.2012.07.018
43. Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. Journal of Educational Psychology, 105(2), 249–265. https://doi.org/10.1037/a0031311
44. Yao, D. (2026). The impact of gamified learning in mathematics education on students' cognitive and affective outcomes: Insights from a second-order meta-analysis. British Educational Research Journal. Advance online publication. https://doi.org/10.1002/berj.70144

APPENDIX

A. Full database search strategies

The searches were executed on 7 August 2026. The complete search strings are reproduced below as run; database field codes and limits are retained to support reproducibility. Search execution, retrieval counts, eligibility criteria, and record-management decisions are reported in the Method and Results sections and are not duplicated here.

A1. Scopus

Exact search syntax executed on 7 August 2026. Field codes and database-specific limits are retained below.

```
TITLE-ABS-KEY(
  (
    "game-based learning" OR "game based learning" OR
    "digital game-based learning" OR "digital game based learning" OR DGBL OR
    "game-based instruction" OR "game based instruction" OR
    "game-based teaching" OR "game based teaching" OR
    "educational game*" OR "learning game*" OR
    "serious game*" OR "digital learning game*" OR
    "digital game*" OR "computer game*" OR "video game*" OR
    "mobile game*" OR "simulation game*" OR
    "computer-based game*" OR "computer based game*" OR
    "mathematics game*" OR "math game*" OR
    "board game*" OR "card game*" OR "tabletop game*" OR
    "escape room*"
  )
  AND
  (
    mathematic* OR math OR maths OR numeracy OR arithmetic OR
    algebra* OR geometr* OR fraction* OR decimal* OR
    "number sense" OR "word problem*" OR
    "mathematical problem solving" OR "mathematical reasoning"
  )
  AND
  (
    "K-12" OR K12 OR kindergarten* OR
    "primary school*" OR "elementary school*" OR
    "middle school*" OR "junior high*" OR
```

```
   "secondary school*" OR "high school*" OR
   "school student*" OR grader* OR pupil* OR child*
  )
  AND
  (
   achiev* OR perform* OR "learning outcome*" OR
   "learning gain*" OR test* OR score* OR
   knowledge OR skill* OR effect* OR evaluat*
  )
  AND
  (
   experiment* OR "quasi-experiment*" OR "quasi experiment*" OR
   random* OR control* OR compar* OR intervention* OR trial
  )
)
```

A2. Web of Science Core Collection

Exact search syntax executed on 7 August 2026. Field codes and database-specific limits are retained below.

```
TS=(
  "game-based learning" OR "game based learning" OR
  "digital game-based learning" OR "digital game based learning" OR DGBL OR
  "game-based instruction" OR "game based instruction" OR
  "game-based teaching" OR "game based teaching" OR
  "educational game*" OR "learning game*" OR
  "serious game*" OR "digital learning game*" OR
  "digital game*" OR "computer game*" OR "video game*" OR
  "mobile game*" OR "simulation game*" OR
  "computer-based game*" OR "computer based game*" OR
  "mathematics game*" OR "math game*" OR
  "board game*" OR "card game*" OR "tabletop game*" OR
  "escape room*"
)
AND TS=(
  mathematic* OR math OR maths OR numeracy OR arithmetic OR
  algebra* OR geometr* OR fraction* OR decimal* OR
  "number sense" OR "word problem*" OR
  "mathematical problem solving" OR "mathematical reasoning"
)
AND TS=(
  "K-12" OR K12 OR kindergarten* OR
  "primary school*" OR "elementary school*" OR
  "middle school*" OR "junior high*" OR
  "secondary school*" OR "high school*" OR
  "school student*" OR grader* OR pupil* OR child*
)
AND TS=(
  achiev* OR perform* OR "learning outcome*" OR
  "learning gain*" OR test* OR score* OR
  knowledge OR skill* OR effect* OR evaluat*
)
AND TS=(
```

```
  experiment* OR "quasi-experiment*" OR "quasi experiment*" OR
  random* OR control* OR compar* OR intervention* OR trial
)
```

A3. ERIC

Exact search syntax executed on 7 August 2026. Field codes and database-specific limits are retained below.

```
(
  descriptor:"Game Based Learning" OR
  descriptor:"Educational Games" OR
  "game based learning" OR
  "digital game based learning" OR
  "serious game" OR "serious games" OR
  "mathematics game" OR "mathematics games"
)
AND
(
  descriptor:"Mathematics Education" OR
  descriptor:"Mathematics Instruction" OR
  descriptor:"Mathematics Achievement" OR
  descriptor:"Elementary School Mathematics" OR
  descriptor:"Middle School Mathematics" OR
  descriptor:"Secondary School Mathematics" OR
  mathematics OR math OR numeracy
)
AND
(
  descriptor:"Instructional Effectiveness" OR
  descriptor:"Program Effectiveness" OR
  descriptor:"Comparative Analysis" OR
  descriptor:"Academic Achievement" OR
  achievement OR performance OR effectiveness OR
  experimental OR experiment OR "control group"
)
AND
(
  educationallevel:"Kindergarten" OR
  educationallevel:"Elementary Education" OR
  educationallevel:"Middle Schools" OR
  educationallevel:"Secondary Education" OR
  educationallevel:"High Schools"
)
```

A4. ProQuest Education Database

Exact search syntax executed on 7 August 2026. Field codes and database-specific limits are retained below.

```
NOFT(
  "game-based learning" OR "game based learning" OR
  "digital game-based learning" OR "digital game based learning" OR DGBL OR
  "game-based instruction" OR "game based instruction" OR
```

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VII July 2026

```
  "game-based teaching" OR "game based teaching" OR
  "educational game*" OR "learning game*" OR
  "serious game*" OR "digital learning game*" OR
  "computer-based game*" OR "computer based game*" OR
  "mathematics game*" OR "math game*" OR
  "board game*" OR "card game*" OR "tabletop game*" OR
  "escape room*"
)
AND NOFT(
  mathematic* OR math OR maths OR numeracy OR arithmetic OR
  algebra* OR geometr* OR fraction* OR decimal* OR
  "number sense" OR "word problem*" OR
  "mathematical problem solving" OR "mathematical reasoning"
)
AND NOFT(
  "K-12" OR K12 OR kindergarten* OR
  "primary school*" OR "elementary school*" OR
  "middle school*" OR "junior high*" OR
  "secondary school*" OR "high school*" OR
  "school student*" OR grader* OR pupil*
)
AND NOFT(
  achiev* OR perform* OR "learning outcome*" OR
  "learning gain*" OR test* OR score* OR
  knowledge OR skill* OR effect* OR evaluat*
)
AND NOFT(
  experiment* OR "quasi-experiment*" OR "quasi experiment*" OR
  random* OR control* OR compar* OR intervention* OR trial
)
```