Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

**Received:** 11/10/2026 - **Accepted:** 03/11/2026
**Vol**: 01.06 – November
**DOI**: 10.66104/hpmntf82
**Pages:** 1-39

# Artificial Intelligence in Educational Games for Personalized Learning Adapting Activities Pace and Feedback to Student Difficulties

Vinicius Gomes de Oliveira
Faculdade Presidente Antônio Carlos, Brazil
Corresponding author: vg064444@gmail.com
ORCID ID: https://orcid.org/0009-0007-0151-3296

## Abstract

Artificial intelligence can enable educational games to select activities, regulate progression, and deliver feedback in response to students' difficulties. Successful assisted gameplay, however, does not establish independent mastery. This critical narrative and thematic review examines how adaptation should be interpreted and evaluated when independently demonstrated learning is the primary endpoint. Targeted web discovery documented in the original manuscript was supplemented through October 31, 2026. The selective corpus comprises 28 sources, including 11 game-specific empirical publications and three adjacent empirical studies; 22 sources (78.6%) date from 2021–2026. A contemporary systematic review provides broader field coverage, whereas the present contribution concerns independent mastery, over-assistance, uncertain learner-state inference, teacher supervision, and prospective evaluation. Selected primary studies show favorable, mixed, and null findings under different instructional conditions. Increased correctness need not indicate misconception repair, and strong fixed support can limit the incremental benefit of adaptation. Activities, pace, and feedback are therefore treated as distinct instructional decisions. An author-developed, unvalidated adaptation cycle and proposed metrics specify assistance exposure, severity of feedback errors, independent assessment, retention, transfer, implementation fidelity, and educational opportunity. No new empirical experiment or pooled effect estimate is reported. The proposed adoption

criterion is retained and transferable learning after unnecessary assistance is withdrawn, alongside acceptable workload, accessibility, privacy, and student agency.

**Keywords:** artificial intelligence; educational games; personalized learning; adaptive feedback; learning assessment

## 1. Introduction

Educational games can place learners in situations where an action has an observable consequence: a fraction moves a character along a number line, a programming decision changes a simulated environment, or a scientific explanation unlocks a new investigation. Students nevertheless arrive with different prerequisite knowledge, reading proficiency, strategies, and support needs. A fixed sequence can therefore provide excessive repetition for one student while confronting another with tasks whose conceptual prerequisites are missing. The educational problem is not simply to make a game easier or more entertaining. It is to determine which challenge and support are appropriate for a particular learning objective at a particular moment.

Artificial intelligence (AI) expands the available means of responding to those differences. A student model can estimate knowledge from sequences of responses, an instructional policy can choose a subsequent task, and a language model can formulate a hint. These functions should be distinguished because estimating a learner's state, selecting an intervention, and expressing feedback involve different kinds of evidence and failure. Knowledge-tracing research supplies methods for representing changing competence, but a model that predicts the next answer accurately has not thereby shown that its recommendations cause learning (Shen et al., 2024).

Recent game-specific evidence illustrates both opportunity and limitation. A randomized science-learning study found an advantage for a ChatGPT-supported game condition incorporating examples over a game-only condition. Conversely, an investigation of adaptive feedback in a fraction game reported improved response correctness without a reduction in the targeted misconception errors. These findings motivate a more precise question than whether AI makes educational games effective: which adaptation changes which learning process, for whom, compared with what alternative, and with what consequences after assistance is withdrawn (Chen & Chang, 2024; Demedts et al., 2025)?

The question is particularly important when game progress becomes a proxy for educational progress. A student can finish a mission by following increasingly explicit hints without acquiring the ability to solve an unfamiliar problem. Another student may take longer because they are explaining a strategy, using an accessibility feature, or

testing an alternative representation. A system optimized for completion speed could reward dependence and classify productive reflection as difficulty. Educational interpretation requires an account of what the learner knows, what the interface demands, and what help was supplied.

Kaşarcı and Yurt (2026) provide a broader systematic account of contemporary AI-supported game-based learning. The present review has a narrower purpose: interpreting and evaluating adaptation when independent mastery, rather than assisted gameplay performance, is the endpoint. Its novelty lies in connecting assistance exposure, diagnostic uncertainty, teacher authority, and prospective evaluation; it does not claim comprehensive mapping of the field.

This review asks how AI-enabled educational games can adapt activities, pace, and feedback to students' difficulties while supporting independently demonstrated learning. Its subsidiary questions concern the reliability of learner-state inference, the comparative evidence for adaptive support, the preservation of agency and access, and the design of credible evaluation. The contribution is a critical synthesis and a proposed teacher-supervised adaptation cycle. The article reports no new intervention, learner dataset, validated diagnostic instrument, or pooled effect estimate. Its recommendations are intended to be testable and locally adaptable rather than universal prescriptions.

## 2. Literature Review

### 2.1 Educational games gamification and personalization

A digital educational game integrates learning objectives into an interactive system of goals, rules, challenges, and consequences. Gamification applies selected game elements to an activity that need not constitute a complete game. This distinction matters when interpreting evidence: an adaptive quiz with points, a narrative programming game, and a medical simulation do not represent the same intervention. The broader gamification literature examines cognitive, motivational, and behavioral outcomes separately and reports variation across implementations; its findings cannot establish an incremental benefit from AI in educational games (Sailer & Homner, 2020).

Personalization is used here as the deliberate adjustment of instruction to a learner's current needs and goals. Adaptivity refers to system-initiated adjustment based on

observed or inferred information. Adaptability refers to adjustments selected by the learner or teacher. These mechanisms can coexist: a game may recommend a prerequisite mission while allowing a teacher to override it or a learner to select a representation. Tailoring decorative rewards or narrative themes may influence participation, but it is analytically different from tailoring the mathematical operation or reasoning step being practiced. Research on tailored gamification supports keeping the tailoring approach and the adapted game element explicit (Hong et al., 2024).

The present scope includes rule-based adaptation, probabilistic learner models, machine-learning methods, and generative language assistance when they inform the game-related instructional cycle. Not every responsive game requires machine learning, and not every adaptive rule warrants an AI label. A transparent rule selecting a misconception-specific example may be more suitable than a complex model when the relevant decision is narrow and its basis is well understood. Conversely, generation can vary wording without adapting the underlying pedagogical strategy. Claims about intelligence should therefore identify the mechanism rather than rely on product terminology.

### 2.2 Formative feedback and instructional scaffolding

Formative feedback provides information intended to improve subsequent thinking or action. It can identify an error, explain its basis, suggest a strategy, or prompt monitoring. The relevant distinction is between displaying information and enabling a learner to use it. A response such as “incorrect” marks an outcome but gives little guidance about the representation or procedure that caused it. An explanation that reveals the entire solution can also be educationally inadequate if it removes the reasoning opportunity (Shute, 2008).

Hattie and Timperley (2007) distinguish feedback concerning the task, the process, and self-regulation, while cautioning that the effects of feedback depend on its content and use. For an educational game, this suggests separating a statement about the current answer from guidance about how to inspect a strategy and from a prompt to evaluate one's own understanding. The same explanatory message may be useful after an attempted solution and disruptive before the student has had a reasonable opportunity to act.

Scaffolding is temporary instructional support intended to enable activity that would otherwise be difficult. Its eventual withdrawal is central to the learning objective. A meta-analysis of 49 studies and 154 effect sizes reported a positive average effect of scaffolding in digital game-based learning, with substantial heterogeneity. This supports the relevance of instructional support but does not isolate AI, prove that learner-specific timing is superior, or establish that all scaffolds benefit all groups (Cai et al., 2022).

### 2.3 Learner models and uncertainty

Knowledge tracing represents learning as a changing state inferred from performance over time. The foundational work of Corbett and Anderson (1994) modeled the probability that particular procedural rules had been learned and used those estimates to select practice. Contemporary research includes a wider range of model families and assumptions (Shen et al., 2024). For games, the attraction is that a sequence of actions can inform a more differentiated account of difficulty than a single total score.

Nevertheless, observations are not transparent windows into understanding. A wrong answer can reflect a misconception, an input error, unfamiliar vocabulary, or an accessibility barrier. A correct answer can reflect guessing, imitation, or the use of a revealing hint. The learner model should therefore retain the conditions under which a response was produced. A probability of success conditional on extensive assistance should not be interpreted as the probability of independent mastery.

For this review, uncertainty is an operational requirement rather than a residual label. When several explanations remain plausible, the system can select a low-stakes diagnostic task, ask a brief clarification question, or request teacher review. It should not convert an ambiguous interaction into a durable deficit label. The instructional consequence of an uncertain estimate can be more consequential than the estimate's numerical precision, especially when it determines access to advanced content.

### 2.4 Motivation agency and accessibility

Self-determination theory relates educational motivation to support for autonomy, competence, and relatedness. Adaptive assistance can support competence by making challenges intelligible, yet tightly controlled progression can undermine ownership of learning. More rewards and longer playtime therefore cannot be assumed

to represent better motivation. The quality of participation and the learner's reasons for engaging require separate attention (Ryan & Deci, 2020).

Accessibility also differs from inferring an individual learning preference. The Universal Design for Learning Guidelines emphasize options for engagement, representation, and action or expression. Such options can be available without requiring the system to diagnose a disability or assign a fixed learner type (CAST, 2024). In the proposed interpretation, personalization should complement accessible design rather than make access conditional on successful profiling. Students should be able to obtain readable text, alternative input, or additional processing time without first being classified as struggling.

## 3. Methodology

### 3.1 Review design and scope

This study is a critical narrative and thematic review with a selective evidence map. It examines instructional mechanisms and interpretive boundaries rather than estimating a population-wide average effect. The review does not claim systematic-review completeness, PRISMA compliance, or independent duplicate screening. Literature discovery, source checking, and synthesis were conducted in a single AI-assisted workflow. There was no second independent screener, extractor, or quality assessor; consequently, no inter-rater agreement statistic is reported.

The original manuscript documented discovery and verification through October 28, 2026. A targeted revision update was conducted through October 31, 2026, including the review identified by the evaluators and selected primary sources from its inventory. This update is not an exhaustive rerun of the field search. Priority was given to publications from 2021–2026, with older sources retained for established concepts. English-language materials were used. Publication year follows the final issue when available; advance online publication is identified when issue information was not established.

### 3.2 Discovery routes and eligibility

Targeted public web searches were used to discover scholarly records and locate publisher pages, university or research-institute repositories, and official guidance. Representative query combinations included “adaptive educational games artificial

intelligence personalization learning feedback 2025 2026," "adaptive digital game based learning meta analysis 2024 2025 personalization," "adaptive educational game explanatory feedback," and combinations of "knowledge tracing," "game-based learning," "generative AI," "scaffolding," "pace," and "self-regulated learning." Follow-up searches used exact titles, authors, and digital object identifiers to verify relevant records.

These were web-discovery queries, not searches executed inside subscription bibliographic databases. Scopus, Web of Science, ACM Digital Library, and IEEE Xplore were not directly queried as database search interfaces. Publisher platforms and repositories served as routes to source content and metadata verification. ERIC and PubMed records were used for discovery and corroboration, not as a basis for claiming exhaustive searches of those indexes. Wikipedia, commercial promotional summaries, and unverified general-interest accounts were not included in the analytical corpus.

Eligible empirical publications examined educational games with adaptive support, AI assistance, or a directly relevant feedback mechanism. Adjacent nongame studies were retained only when they tested a mechanism central to the argument, such as generative tutoring, feedback, or dependence on assistance. Reviews and theoretical sources informed interpretation; official guidance informed governance and accessibility. Broad AI-adoption commentary, purely recreational difficulty adjustment without an educational endpoint, and unsupported vendor claims were not retained as effectiveness evidence. References within relevant records informed targeted follow-up discovery.

The revision update used exact-title and DOI queries for Kaşarcı and Yurt (2026), Huang and Chen (2025), Panjaburee et al. (2025), and Gupta et al. (2024). Candidate selection targeted instructional component separation, longitudinal personalization, and assessment fairness. Sources rated favorably in the broader review were prioritized where relevant, but their primary designs were checked rather than adopting the review's quality labels as this article's own appraisal. Chen and Chang (2024), already included, was retained; Huang and Chen (2025) and Panjaburee et al. (2025) were added. Gupta et al. (2024) supplies complementary predictive-model evidence. These additions do not imply that all other omitted studies were ineligible or inferior.

### 3.3 Corpus composition and access limits

The revised corpus contains 28 sources: 11 game-specific empirical publications; three adjacent empirical studies outside games; six research syntheses or technical surveys; four foundational theoretical or methodological sources; and four official guidance documents. Twenty-two sources (22/28 = 78.6%) were published in 2021–2026; 24 (24/28 = 85.7%) date from 2020–2026. Categories reflect each source's primary interpretive role, not interchangeable evidence strength. Table 1 inventories the complete selected corpus. Publications included within a cited systematic review are not counted as individually reviewed primary sources unless separately incorporated here.

The 11 game-specific publications are Adam et al. (2026), Aydemir and Tüzün (2026), Chen and Chang (2024), Demedts et al. (2024, 2025), Faber et al. (2024), Liu and Hwang (2024), Wang et al. (2025), Huang and Chen (2025), Panjaburee et al. (2025), and Gupta et al. (2024). Their inclusion does not imply that all evaluate machine-learning personalization: structured adaptive and metalinguistic feedback supply component evidence, whereas the generative studies address additional mechanisms. Publication counts must not be interpreted as counts of independent datasets; independence between related publications was not established.

Accessible full texts were examined where available. For subscription-restricted material, analysis was limited to publisher-provided abstracts, visible methodological information, and verified bibliographic records. Consequently, detailed risk-of-bias judgments, unavailable subgroup estimates, and unreported implementation parameters were not reconstructed. A complete deduplicated ledger of every web result screened was not maintained, so initial identification and exclusion counts are not reported. The final inventory is auditable, but the process cannot support an exhaustive-coverage claim.

### 3.4 Extraction and thematic analysis

Source information was organized by educational setting, study design, learner population, adaptation target, comparator, measured outcome, principal finding, access level, and inference limit. Empirical support was interpreted according to whether the design could distinguish adaptation from additional content, additional attention, or a different interface. Particular attention was given to independent testing,

delayed assessment, comparison quality, and the difference between a statistically nonsignificant result and evidence of equivalence.

Source-level descriptors distinguish randomized comparisons, factorial experiments, quasi-experiments, longitudinal associational analyses, design-based research, and predictive-model studies. Adjacent tutoring trials remain separate from game evidence. These descriptors identify inferential boundaries; they are not a formal risk-of-bias score or an independently replicated quality assessment. Access labels in Table 1 preserve the original manuscript's documented access and identify the revision update's additional publisher or repository checks. No second independent screener or extractor was introduced.

Synthesis used three adaptation targets: activities, pace, and feedback. Cross-cutting themes were diagnostic uncertainty, learner agency, accessibility, teacher mediation, and evaluation validity. These are analytical categories developed for this article, not formal evidence grades. Numerical findings are reported only where verified in the retrieved source record. No new effect size was calculated, no incompatible outcome percentages were pooled, and no causal conclusion was drawn from the proposed architecture. The adaptation cycle and research protocol in the Discussion are original conceptual recommendations requiring prospective validation.

## 4. Results

### 4.1 Evidence supports specific mechanisms rather than universal superiority

#### 4.1.1 Positioning the selected evidence within the broader field

Kaşarcı and Yurt (2026) synthesized 55 empirical studies published in 2021–2026. Their appraisal classified only four as high quality, with most rated moderate or low. They discuss adaptive difficulty, LLM scaffolding, intelligent NPCs, and stealth assessment, while identifying over-reliance, equity, and longitudinal gaps. This breadth corrects any impression that the current selection exhausts the field. Its favorable overall interpretation remains conditional on heterogeneous designs and comparators; it does not establish that every adaptive mechanism improves independent learning. The present framework develops evaluation requirements rather than reproducing that evidence map.

The selected publications support a conditional account of adaptive games. Some report advantages for learning or interaction, but the technologies, comparators, age groups, and outcomes differ considerably. Older game-based learning synthesis establishes that design and instructional support matter, without establishing the added effect of present-day AI (Clark et al., 2016). The relevant comparison for personalization is therefore not simply a game against no game, but an adaptive version against a sufficiently strong alternative with comparable content and instructional opportunity.

Chen and Chang (2024) randomly assigned 202 seventh graders to game-only, GameGPT, or GameGPT-with-examples conditions. The condition incorporating examples outperformed game-only on achievement, and the AI-supported groups reported greater perceived competence. The comparison is useful because the game context was retained. However, the contribution of examples and the contribution of generative interaction should not be collapsed into a single general AI effect. The verified record supports a bounded instructional advantage rather than an inference about all subjects or long-term independent performance.

Wang et al. (2025) compared a large-language-model adaptive contextual game with a conventional contextual game among fifth graders using a quasi-experimental design. The publisher abstract reports advantages in achievement and flow, together with lower cognitive load and favorable behavior patterns. This adds game-specific evidence for contextual assistance, but the design should not be represented as randomized. Parameters not available in the retrieved record, including a verified sample count and detailed delayed outcomes, are not supplied here.

The available studies also resist a uniformly positive account. Aydemir and Tüzün (2026) developed a C++ role-playing game through five design-based research cycles involving 70 participants. The system combined retrieval-augmented generation and language models. Short-term attitude and motivation changes were not statistically significant, although qualitative and log evidence suggested more analytical interaction after technical stabilization. These results support design refinement and behavioral hypotheses; they do not demonstrate a stable causal achievement effect. Table 2 summarizes the principal game-specific comparisons and their boundaries.

Huang and Chen (2025) used a 2 × 2 experimental design with 160 seventh graders to distinguish instructional video from a generative chatbot. The reported science-learning advantage was associated with video availability; combining tools also affected extraneous cognitive load. This component comparison cautions against assigning all benefits of a combined intervention to generative adaptation. The publisher abstract does not establish delayed independent mastery. Panjaburee et al. (2025) followed 372 lower secondary students using fuzzy-logic and decision-tree personalization in digital citizenship learning. Their repeated-measures PLS-SEM analysis examined perceived sustainability and motivational pathways. It extends the temporal perspective but does not by itself establish a randomized adaptation effect or delayed independent mastery. Its longitudinal associations should not be relabeled as causal proof.

### 4.2 Adaptive explanatory feedback does not guarantee conceptual repair

Demedts et al. (2024) studied 288 fourth graders assigned to corrective-feedback, explanatory-adaptive-feedback, or control conditions. Near-transfer fraction knowledge improved in both game conditions. The explanatory condition showed reduced mathematics anxiety, while motivational and self-concept findings were mixed. The pattern is important because cognitive and affective outcomes did not move together. It does not justify claiming that explanatory adaptation was uniformly superior across all outcomes or that successful fraction practice establishes broader mathematical competence.

A later publication examined item-level responses from 197 fourth graders assigned to corrective or explanatory adaptive feedback. Correctness improved more strongly with explanatory adaptive feedback, yet neither condition reduced the targeted misconception errors. Prior knowledge and misconception type did not moderate the reported effect (Demedts et al., 2025). The distinction between solving more items and changing the underlying misconception should remain visible in evaluation. Repeated successful play may be compatible with a strategy that is locally effective but conceptually incomplete.

The 2024 and 2025 Demedts publications concern related fraction-game feedback interventions. The accessible publisher and institutional records reviewed for this revision did not establish whether the samples overlap. They are therefore treated as

two publications with potentially related data, not two demonstrably independent replications. Their participant counts are not summed to estimate a unique learner total. Full-text confirmation or author-provided provenance would be needed to resolve sample independence.

Liu and Hwang (2024) investigated metalinguistic corrective feedback with 76 university learners of English. Feedback referring to grammatical rules was compared with ordinary immediate correction within digital games. The reported results favored grammar learning and identified benefits for learners with lower initial motivation and self-efficacy. This is evidence about the instructional form of feedback, not proof that a generative model is necessary. It suggests that a well-designed structured explanation is an important comparator for any claim that AI-generated feedback adds value.

Across these examples, the form of the response is at least as important as its apparent personalization. An explanation must address the reasoning difficulty rather than merely vary the language of encouragement. Evidence that students notice a hint, change their action, or answer the next item correctly is useful process information. Demonstrating conceptual repair requires additional tasks that expose the original error in a new representation, ideally after the explanatory support is removed.

### 4.3 Adaptive scaffolding requires a discriminating comparator

Faber et al. (2024) randomized 62 medical students in a matched-pairs yoked-control study involving six emergency-game scenarios. Adaptive and nonadaptive scaffolding did not differ detectably on the principal comparisons, test performance, or engagement. A substantial proportion of control scenarios nevertheless received support that happened to fit their needs. Exploratory findings about tailored support cannot replace the primary randomized result. This example shows why a nominally nonadaptive comparator may still provide instruction appropriate to many learners.

The result has two implications for interpretation. First, adaptation should not be declared ineffective merely because a small study did not detect a difference. Second, the result cannot be converted into proof that adaptation works by emphasizing favorable exploratory contrasts. Researchers need to describe how often the intervention actually differs from the comparison and whether the alternative already contains strong scaffolds. Educational necessity and technological novelty are separate questions.

Pacing is particularly difficult to isolate in the selected corpus. A system can change scenario order, offer additional practice, or deliver feedback after a pause, but such changes do not constitute a dedicated test of adaptive temporal scheduling. The present evidence map therefore provides a stronger basis for discussing feedback and supported interaction than for estimating the independent effect of adaptive pacing. Claims that AI determines an optimal learning rhythm for every student exceed what this selective corpus can establish.

### 4.4 Learners differ in how they use the same assistance

Adam et al. (2026) examined generative-AI-supported self-regulated science gameplay with 48 junior high school students categorized by prior achievement. Higher-achieving students used assistance more strategically, whereas lower-achieving students more frequently sought basic information and formulas. The study did not establish that achievement grouping caused these differences, nor that an adaptive remedy had already closed the gap. It provides a recent warning that equal access to a conversational tool does not ensure equally productive use.

A game can consequently appear personalized while placing most of the burden of effective help-seeking on the learner. Students who already know how to formulate a diagnostic question may obtain useful guidance, while students unable to identify their misconception repeatedly request an answer. A support policy should therefore examine the sequence after a hint: whether the student revises a strategy, explains a step, or returns to the same difficulty. Frequency of help requests alone is ambiguous and should not be treated as evidence of either engagement or dependence.

The instructional response should remain provisional. A pattern of repeated requests may indicate an inaccessible prompt, a prerequisite gap, a poorly written hint, or uncertainty about the game interface. Labeling the learner as weak without checking these alternatives shifts responsibility from design to the student. The proposed cycle accordingly treats learner-state inference as revisable and records interface failures separately from conceptual errors.

Gupta et al. (2024) evaluated deep-learning stealth assessment using gameplay and written reflection from 119 middle school learners in Crystal Island. The study examined prediction and encoded bias with respect to gender and prior gaming experience. It adds evidence about assessment-model fairness, not a randomized

test that prediction-driven adaptation causes learning. A model may predict outcomes while still recommending ineffective instruction; predictive discrimination, calibration, subgroup error, and educational consequences require separate evaluation.

### 4.5 Adjacent AI studies clarify benefits and dependence risks

Three nongame studies provide mechanism-relevant evidence, but none estimates an educational-game effect. Kestin et al. (2025) conducted a randomized comparison with 194 undergraduate physics students and reported greater learning in less time with a carefully designed AI tutor than with the comparison active-learning instruction. The intervention was a structured pedagogical system rather than unrestricted conversation. The finding supports investigation of content-aligned guidance, while its setting, duration, and comparison constrain transfer to younger learners and game environments.

Bastani et al. (2025) studied nearly 1,000 high school mathematics students using two forms of generative assistance. Unrestricted assistance improved performance during practice but was followed by poorer performance when assistance was removed; a guarded tutor mitigated that harm. This distinction directly motivates assessment beyond gameplay, but the result should not be reported as a trial of educational games or as proof that a guarded tutor necessarily improves independent learning. Preventing a negative effect and producing a positive learning gain are different achievements.

Bauer et al. (2025) compared AI-generated adaptive feedback with static feedback in a higher-education statistics environment. The reported absence of a significant task-performance advantage and less favorable interest findings challenge the assumption that generation necessarily improves a structured exercise. As adjacent evidence, it strengthens the case for a strong static-feedback comparator and for measuring autonomy-related experience separately from performance. It does not negate positive findings in different tasks.

Taken together, these results justify neither blanket adoption nor blanket rejection. They indicate that assistance design, comparison quality, and assessment conditions determine the meaning of an apparent benefit. A teacher-approved explanatory message, a constrained tutor, and a general-purpose chatbot should not be combined under one intervention label. Likewise, improved confidence, reduced burden,

immediate success, and durable competence should be reported as separate outcomes.

## 5. Discussion

### 5.1 A teacher-supervised adaptation cycle

The following cycle is proposed as an original design and evaluation framework, not a validated scoring instrument. Its organizing sequence is to define the learning objective, observe relevant evidence, estimate the learner's state with uncertainty, select an authorized instructional action, deliver that action, and check subsequent independent performance. Teacher authority and learner access span the entire cycle. Table 3 translates these stages into observable decisions and records.

The first stage is to specify what counts as learning before selecting an algorithm. An objective such as comparing fraction magnitudes requires different evidence from an objective such as completing a set of number-line placements. The designer should identify prerequisite concepts, likely errors, permitted representations, and acceptable support. This instructional specification makes the adaptation interpretable: the system can explain why a selected task is relevant instead of merely announcing that an algorithm recommended it.

The second stage is to separate observations from interpretations. The record can state that a learner selected a particular distractor after reading a hint; it should not state as an observed fact that the learner has a stable misconception. An interpretive layer may propose that explanation, together with alternatives and confidence. The distinction permits correction when a teacher or learner identifies an interface problem. It also prevents game telemetry from acquiring unwarranted diagnostic authority.

For probabilistic learner models, confidence must be distinguished from correctness. Prospective validation should assess calibration using a calibration curve and, where suitable, a Brier score or calibration-error measure, with the prediction target and assistance conditions stated. A confident but miscalibrated model must not acquire diagnostic authority. Competing interpretations, teacher corrections, and student appeals should be retained as revision records rather than overwritten by the next automated prediction.

The third stage is to select from a bounded action set. Permitted actions might include another representation, a prerequisite task, a reduced language load, an optional worked example, or teacher referral. The system should not invent an unrestricted curriculum pathway simply because it can generate content. Every action must preserve the intended construct unless the teacher deliberately changes the objective. A decision to provide text-to-speech, for example, may preserve the mathematics objective, whereas eliminating the fraction comparison changes it.

The final stage is to test the consequence of the adaptation. A useful hint is one that enables a subsequent meaningful attempt, but the cycle remains incomplete until later evidence checks whether support can be reduced. This check can occur within the game and through independent tasks outside it. The framework does not assume that every adaptation will succeed. It requires failed adaptations, overrides, and unresolved difficulties to remain visible so that the instructional policy can be revised.

### 5.2 Adapting activities to specific difficulties

Activity adaptation should target an interpretable obstacle. Consider an illustrative fraction game in which a learner repeatedly chooses the fraction with the larger denominator as the larger quantity. Rather than immediately increasing or decreasing a global difficulty level, the system can propose a comparison using equal-sized visual units, then request a prediction before showing the result. A later item can remove the visual support and change the surface context. This is a hypothetical design example, not an intervention evaluated in the reviewed studies.

The distinction between prerequisite repair and reduced expectations is essential. Recommending a simpler representation can create an entry point while retaining the original goal. Permanently assigning easier problems can instead restrict opportunity. A teacher dashboard should therefore display the objective being pursued, the basis for a detour, and the condition for returning to the shared progression. Students who demonstrate readiness should be able to exit a remedial branch without completing unnecessary repetition.

Activities should vary along identifiable dimensions. Numerical magnitude, number of reasoning steps, language complexity, representation, distractor structure, and required explanation are different sources of challenge. Raising several dimensions at once makes an error difficult to interpret. During diagnosis, changing one dimension

while holding others reasonably stable can help distinguish a mathematical difficulty from a reading difficulty. During transfer practice, deliberately varying several dimensions may be appropriate, provided that assessment recognizes the additional demands.

A content bank can combine teacher-authored tasks with generated variants, but generated novelty requires verification. A mathematically valid problem may have an ambiguous diagram, multiple defensible answers, or vocabulary inappropriate for the learner. Before release, generation should be constrained by objective tags, prerequisite relations, answer validation, and age-appropriate language requirements. Where reliable validation is unavailable, the system should use reviewed material or refer the case to a teacher rather than expose the student to an unverified task.

The goal is not to select the most individually distinctive activity. It is to select an activity whose educational purpose can be explained and assessed. Two learners may appropriately receive the same task for different reasons, and a student may benefit from a nonpersonalized whole-class discussion. Personalization is useful when it improves access to the objective; it is not an obligation to make every learning experience different.

### 5.3 Adapting pace without confusing speed and mastery

Pace includes time allowed, spacing of practice, frequency of review, duration of a session, and the conditions for moving forward. These dimensions should not be collapsed into response latency. A game can adjust temporal pressure independently of conceptual complexity. For a learner using an alternative input device, removing a countdown may improve access without changing what is being learned. A slower response should not automatically trigger remedial content.

The proposed policy first checks whether timing information is interpretable. Pauses caused by connectivity, instructions, interruptions, or assistive technology should be flagged separately where feasible. A learner can also pause voluntarily to plan. The game should avoid equating rapid repeated guessing with fluent knowledge. Accuracy, assistance level, explanation, and recurrence after a delay provide a more educationally meaningful basis for progression than speed alone.

Mastery decisions should specify the evidence required rather than use an opaque completion percentage. A local pilot might require successful independent

performance across varied examples and a later review before advancing. No universal numerical threshold is prescribed here, because the appropriate standard depends on the construct, assessment reliability, and consequences of error. Repeated success on nearly identical items should not substitute for success on new representations.

Pacing adaptation should also avoid endless remediation. A student who repeatedly fails after several different supports may need teacher intervention rather than another generated version of the same exercise. The system should record the number and type of unsuccessful adaptations and make escalation available without penalty. A learner should be able to request a pause, review, or human explanation even when the algorithm predicts readiness to continue.

Classroom pacing introduces a collective constraint. Individual progression must remain compatible with collaborative activities, shared discussion, and the teacher's curriculum. A game can provide optional extension work for students who finish early while retaining a common conceptual checkpoint. In this arrangement, personalization supports participation in the classroom rather than separating students into permanent algorithmic tracks. The educational endpoint remains readiness to engage with the shared objective.

Pace should be operationalized as distinct dimensions: time allowed per attempt, spacing between practice opportunities, session frequency, session duration, and advancement criteria. A future study should identify which dimension changed and preserve the other instructional conditions as far as feasible. Familiarization with controls should precede diagnostic use of interaction speed. Accessibility-related processing time, network latency, and interface recovery should be coded separately from conceptual response time; none alone establishes difficulty or mastery.

### 5.4 Adapting feedback while preserving productive effort

Feedback adaptation should distinguish what is changed: timing, specificity, modality, explanation depth, or the reasoning strategy addressed. A defensible sequence may begin with a prompt to inspect an important feature, continue with a misconception-specific cue, and offer a worked example when a less explicit intervention is insufficient. This sequence is a proposal, not evidence that a fixed hint ladder is optimal. Students and teachers need the ability to bypass a poorly matched hint.

A short feedback message can invite a new action without supplying the answer. In the illustrative fraction game, the system might ask the student to check whether the compared wholes have the same size. The next task then tests whether the learner uses that distinction. Generic encouragement may remain useful socially, but it should not be recorded as conceptual remediation. Similarly, a long explanation should not be treated as effective merely because it sounds individualized.

Generative feedback introduces a separate verification problem. Retrieval from approved material can help constrain a response, yet the presence of retrieved content does not establish that the explanation is correct, complete, or applicable. A hybrid architecture can use deterministic checks for answers and prerequisite relations while allowing a language model to rephrase an approved explanation. The instructional decision and the wording decision should be logged separately so that failure can be traced to the appropriate component.

Generated tasks and feedback require validation of the problem statement, correct answer, diagrams or other representations, and age-appropriate language. Reviewed content banks and deterministic answer or constraint checks should bound generation where feasible. Logs should identify the model and version, prompt/instruction architecture and version, generation settings such as temperature where available, retrieved RAG sources and their versions, content-bank version, adaptation thresholds, support level, teacher override, and evaluation dates. A retrieval citation is not a correctness guarantee. Any material model, content, or policy change requires proportionate revalidation before prior findings are transferred to the changed system.

Students should know when a conversational character is automated and how to report an incorrect explanation. The interface should allow uncertainty and human referral. If the system cannot determine whether a student's reasoning is valid, it should not invent a misconception to justify a response. For open-ended explanations, teacher review or a request for clarification may be more appropriate than immediate categorical judgment.

Assistance should eventually fade where independence is the objective. The evaluation record must distinguish an independently solved task from one solved after a partial cue, a worked example, or an explicit answer. These categories prevent a

game from reporting higher mastery simply because it supplied more help. They also permit investigation of whether a particular feedback mode benefits some learners while delaying independence for others.

Permitted assistance should be specified before evaluation, for example as no hint, a general cue, a targeted cue, a worked example, or an explicit answer. Each attempt should retain both assistance availability and actual help exposure. Over-assistance occurs when delivered support exceeds the level authorized for that task and assessment purpose, including answer leakage that removes the reasoning opportunity. A successful attempt following an explicit answer must not be coded as independent mastery. Support can be faded gradually, while keeping accessible representations available when these do not supply the assessed knowledge.

### 5.5 Agency accessibility and collaborative learning

Personalized instruction should preserve a student's ability to understand and influence the learning pathway. A brief explanation such as “this activity checks the idea that caused difficulty in the previous mission” makes an intervention more intelligible than an unexplained level reduction. The learner should be able to contest an interpretation, request a different representation, and ask for a teacher. These are proposed interface requirements, not outcomes already demonstrated by the selected studies.

Accessible alternatives should be designed before algorithmic adaptation is activated. Text size, keyboard navigation, captions, audio control, and alternative response modes can be provided as ordinary options. Requiring the system to infer a disability from gameplay adds unnecessary uncertainty and can expose sensitive information. A diagnostic prediction should never become the prerequisite for an accommodation that can be offered directly.

Group activities need separate treatment because an individual model can misattribute collaborative performance. A student may follow a peer's reasoning, contribute an unlogged oral explanation, or provide a strategy that another player enters. A shared score cannot establish each participant's mastery. Cooperative games should therefore combine group-level objectives with individual opportunities to explain, practice, and demonstrate understanding. Adaptation should not remove a student from collaboration solely because their interaction count is low.

Equity evaluation must examine opportunity as well as performance. Students with fewer prior digital skills may need interface practice rather than easier academic tasks. Learners working on older devices may receive slower feedback, shorter sessions, or less reliable voice recognition. These conditions can generate apparent ability differences. A credible implementation records technical access and instructional exposure so that infrastructure failures are not interpreted as student deficits.

### 5.6 Data governance and teacher accountability

Official guidance provides reasons to consider privacy, age appropriateness, and accountability alongside pedagogical design. UNESCO emphasizes human-centered educational use and institutional validation, while UNICEF's 2025 guidance frames AI systems in terms of children's rights. These documents guide governance; they are not experimental evidence that a particular game improves learning (Miao & Holmes, 2023; UNICEF, 2025).

For the proposed cycle, a school should specify which data are necessary for each adaptation. Response correctness, task identity, and assistance level may be sufficient for a narrow learning objective. Continuous audio, facial analysis, or broad personal histories should not be collected merely because the platform supports them. A pseudonymous student identifier does not by itself remove the possibility of identification through linked school records or detailed interaction histories.

Provider handling must be visible in procurement and deployment decisions. Relevant questions include whether learner interactions are retained, used for model training, accessible to administrators, or processed by additional services. The generative-AI risk-management profile offers a cross-sector resource for identifying and managing such risks, but it does not certify an educational game or replace local obligations (Autio et al., 2024). The proposed governance workflow assigns an institutional owner for permissions, retention, deletion, incident response, and model changes.

Institutional governance should also state who controls student data, what the provider retains and for how long, whether interactions may be reused for model training, and how learners or authorized representatives request correction or deletion. Collection should be limited to necessary fields; continuous audio or video requires an educational justification rather than technical availability. Vendor changes, incident handling, and policy revisions should have named institutional

decision owners. This framework proposes governance requirements and does not assert compliance with a particular jurisdiction.

Teacher oversight should involve educational competence and workable review mechanisms. A teacher needs access to the original task, the student's response, the support supplied, and the reason for a recommendation. A dashboard containing only a red risk label cannot support independent judgment. Consequential decisions about grading, placement, disability, or discipline should not be delegated to an unvalidated game profile. The teacher should also be able to correct a profile and see whether that correction propagates through later recommendations.

### 5.7 A prospective evaluation protocol

The following protocol is a research proposal. It has not been implemented or validated in this review. Its central comparison is between an adaptive game and the same game with a strong fixed instructional policy. Content coverage, visual design, access time, reward structure, and teacher support should be held comparable where feasible. A conventional instructional condition may answer a broader adoption question, but it cannot by itself isolate the added value of adaptation.

Where resources permit, a third condition can use teacher-selected adaptation. This helps distinguish the value of personalization from the value of automating it. A yoked condition may also be useful, provided that researchers report how often the support happens to match control learners' needs. The contrast should be specified before data collection: adaptive policy versus fixed policy, generative wording versus reviewed wording, or combined workflow versus usual instruction. A study should not claim to identify all three effects from one bundled comparison.

The primary outcome should be an independently administered assessment aligned with the intended construct. It should include unfamiliar items and record whether AI or game assistance is available. Delayed retention and transfer should be prespecified as separate outcomes, with intervals justified by the curriculum. Gameplay success, motivation, self-efficacy, cognitive load, and satisfaction remain useful secondary outcomes but cannot replace a test of independently demonstrated learning. Table 4 defines proposed measures and their denominators.

Randomization should match the delivery structure. Individual allocation may be appropriate when students use the game independently, whereas classroom allocation

may reduce contamination when teachers and peers share instructional practices. Analyses should account for repeated observations and clustering by learner and classroom. The sample-size calculation should use the chosen primary outcome and allocation level. There is no defensible universal minimum sample size for adaptive-game research.

The statistical analysis plan should distinguish educational importance from statistical significance and, where defensible, prespecify a minimally important learning difference. Report effect estimates with uncertainty intervals, attrition by condition, and missing delayed assessments. Intention-to-treat analysis should follow an explicit missing-data strategy when appropriate; a completer-only analysis requires qualification and sensitivity analysis. Sample-size assumptions should address the actual primary outcome, allocation unit, classroom clustering, repeated observations, and expected attrition rather than use a universal numerical minimum.

Assessment scoring should be independent of intervention delivery where feasible. Two qualified assessors can score a prespecified subset independently before adjudication. The reliability statistic must match the scale, rater structure, and intended use: categorical agreement and continuous-score reliability require different models. The threshold must be justified for the selected statistic and educational stakes; a value such as 0.75 is illustrative where appropriate, not a universal requirement for $\kappa$ or any other index. Report agreement before adjudication, sample size, and uncertainty. Low agreement should prompt rubric revision, sensitivity analysis, and explicit limits on interpretation rather than disappear into consensus.

Development data should be separated from the final evaluation. Student-level separation is necessary when interaction traces are used to tune a learner model; item-family separation helps prevent near-duplicate exercises from inflating transfer results. A later temporal cohort can test whether the policy remains useful after curriculum or software changes. Model selection should not repeatedly optimize performance on a test set later presented as independent evidence.

Reporting should identify the exact model and version, prompts or feedback templates, approved content bank, mastery policy, adaptation thresholds, accessibility settings, and evaluation dates. Generative components require repeated testing because identical inputs can yield different outputs. An institution should retain enough

configuration information to explain a consequential recommendation without retaining unnecessary student content. Material changes to the model or instructional policy should trigger proportionate revalidation.

### 5.8 Operational measures and interpretation rules

Independent mastery (IM) is proposed as N meeting the prespecified standard without assistance divided by N independently assessed. Assisted mastery (AM) is N meeting the standard with the permitted assistance divided by N assessed under that support condition. Both require explicit standards, assessment units, and assistance rules. Report missing assessments separately and preserve the prespecified intention-to-treat analysis. A descriptive difference between AM and IM is interpretable only with attention to comparable learners, tasks, timing, and support conditions; it is not automatically a causal effect of help. The primary causal estimate remains the between-condition effect on the independent outcome, with appropriate uncertainty and baseline adjustment.

Misconception persistence can be defined as the number of independently scored diagnostic opportunities displaying a prespecified misconception divided by the number of eligible scored diagnostic opportunities. Correctness and misconception persistence should both be reported because one can improve without the other. A reduction in errors on repeated game items should not be described as transfer unless the assessment genuinely changes the relevant context, representation, or task demands.

Near transfer, far transfer, and delayed retention answer different questions and should be reported separately. Specify task novelty, representation changes, delay length, scoring criteria, and assistance availability at every checkpoint. Include delayed far-transfer tasks where appropriate to the learning objective. A new surface story with an otherwise identical item is not necessarily evidence of far transfer. Missing delayed outcomes remain missing evidence and cannot be replaced by game completion or immediate assisted scores.

Over-assistance (OA) is the number of reviewed episodes exceeding prespecified permitted support divided by all reviewed episodes. Feedback error (FE) is the number of reviewed episodes with substantive instructional error divided by all reviewed episodes; minor imprecision and pedagogical mismatch should be coded separately.

Record independent coding rules, sampling frame, assistance level, numerators, denominators, and uncertainty. An empty denominator is not estimable, never perfect performance. Code severity as minor instructional issue, substantive instructional error, harmful or discriminatory output, or answer leakage/construct compromise. Categories can coexist, so episode counts and category assignments should both be reported. Serious events must remain visible rather than be offset by favorable average scores.

Effort measures should cover the complete instructional workflow: teacher preparation, content approval, monitoring, dashboard management, verification, explanation of recommendations, AI-error correction, and escalation or follow-up. Student time should separate instructional activity from waiting for model responses, network delays, and technical recovery. Compare equivalent objectives and workflows, and report workload jointly with learning outcomes. Lower direct explanation time does not imply a net saving if verification and correction increase. Do not merge workload, learning, and serious privacy or instructional errors into an unvalidated compensatory score.

Fairness analysis should examine performance and exposure across relevant groups while avoiding unsupported comparisons in very small subgroups. The study should report access failures, dropout, hint availability, advanced-content exposure, and erroneous remediation alongside assessment differences. A subgroup with lower average performance should not automatically receive a lower progression standard. The question is whether the system offers equitable opportunity to meet the intended objective and whether its interpretations are comparably trustworthy.

Opportunity indicators should include exposure to hints and advanced content, accessibility-feature use, device limitations, connection failures, and time available for practice. Subgroup estimates require uncertainty intervals and caution where denominators are small. A lower subgroup score should prompt examination of unequal opportunity and measurement validity before changing expectations. Equal tool access is not evidence of equal benefit, and low interaction counts in collaborative work do not justify an individual deficit diagnosis.

Statistical nonsignificance should not be described as equivalence. If a study aims to show that a cheaper system preserves learning, a noninferiority margin and adequate precision should be defined beforehand. Likewise, zero observed serious errors does

not demonstrate zero residual risk. Reporting confidence intervals, episode counts, and evaluation conditions makes the limits of an assurance claim visible. This is particularly important when a small pilot is used to justify a larger deployment.

### 5.9 Implementation pathway and research priorities

An initial deployment should begin with a narrow objective and a reviewed content bank. During a silent evaluation phase, the system can generate recommendations that teachers inspect without automatically changing students' pathways. This allows analysis of inappropriate remediation, missed difficulties, and teacher disagreement. A subsequent controlled phase can authorize low-consequence adaptations while retaining simple ways to pause or override the system. These stages are proposed safeguards rather than evidence of effectiveness.

The implementation team should define who maintains the curriculum mapping, who approves feedback, who responds to a privacy incident, and who decides whether a software change requires renewed evaluation. Educational responsibility cannot be assigned to a model. A maintenance record should distinguish updates to academic content from updates to the learner model, interface, or provider. Otherwise a positive evaluation may be invoked for a materially different intervention.

Future research should isolate adaptation targets instead of treating personalization as a single feature. Factorial or carefully sequenced studies can compare activity selection, pacing, and feedback while controlling instructional exposure. Longer studies should examine whether support can be withdrawn and whether students improve their own help-seeking and self-monitoring. Comparative studies should also test when a simple rule-based policy performs as well as a more expensive generative architecture.

Portuguese-language and other underrepresented contexts require original evaluation rather than translation alone. Vocabulary, curriculum sequences, classroom practices, dialects, device access, and locally familiar examples can influence both game interaction and model interpretation. For Brazilian implementation, assessment materials should be checked against the applicable curriculum and institutional requirements. This article does not present jurisdiction-specific legal advice or claim that evidence from another language validates local deployment.

### 5.10 Limitations

The review is selective and mechanism-oriented. Targeted web discovery cannot establish complete coverage of educational-game research, and the absence of a full screening ledger prevents reconstruction of an exhaustive selection flow. English-language selection and publicly accessible records may privilege particular publishers, countries, and subjects. No independent duplicate screening or extraction was conducted. These limitations affect completeness and interpretive reliability and should remain explicit.

The revision incorporates the previously omitted broad systematic review and three independently checked primary publications. This repairs a specific coverage gap but does not establish completeness. The broader review's study inventory and quality ratings were used for targeted discovery and contextual interpretation, not adopted as independently verified assessments of every included study. Restricted access limits verification of some primary details, and publication-level counts cannot resolve shared datasets.

Some empirical sources remain accessible only through publisher abstracts or institutional records. No complete formal risk-of-bias assessment, unavailable statistical detail, or comprehensive implementation audit is supplied. The 11 game-specific publications represent heterogeneous intervention and predictive-model designs, not 11 independent causal demonstrations of AI benefit. Related Demedts publications may share participants or infrastructure, and independence remains unresolved. Learner ages, educational levels, subjects, and AI mechanisms differ; the selected findings do not justify unqualified transfer across these contexts. Longitudinal independent-learning evidence and isolated evidence for adaptive pacing remain limited.

The proposed cycle, metrics, and implementation stages are conceptual contributions. They have not been validated as a package, benchmarked against competing frameworks, or shown to improve learning. The recommendations specify what future evidence should address rather than claim that such evidence already exists. Rapid model and interface changes further limit the lifetime of any product-specific conclusion. Durable claims should concern verified instructional mechanisms and outcomes, with system versions and contexts stated explicitly.

## 6. Conclusion

AI can support educational-game personalization, but the selected evidence does not justify universal superiority. Immediate success while support is available must be distinguished from skill acquisition demonstrated after unnecessary help is withdrawn. A broader contemporary review contextualizes the field; the present contribution is a narrower, independent-mastery-centered interpretation of adaptation and a prospective evaluation framework. Favorable, mixed, and null comparisons underline the importance of instructional architecture, assistance constraints, learner characteristics, and comparator quality.

Activities, pace, and feedback should be treated as distinct instructional decisions. A game should identify the objective behind an adaptation, retain uncertainty about the learner's state, preserve accessible alternatives, and test whether support leads toward independent performance. The proposed teacher-supervised cycle makes those requirements explicit while retaining professional authority over consequential decisions. It is a design and research proposal, not a validated instrument or certification.

The most defensible adoption criterion is a demonstrated improvement in independently assessed, retained, and transferable learning under acceptable conditions of effort, fairness, privacy, and student agency. Evaluation should compare strong alternatives, report denominators and uncertainty, and preserve negative findings. Educational personalization succeeds when students become more capable of learning and acting without unnecessary assistance, not when the system merely becomes more capable of guiding them through a game.

## Data Availability

No new participant-level data were collected or generated. The source inventory is provided in Table 1, and the cited publications and official documents are identified in the reference list. No proprietary learner dataset is associated with this review.

## Ethical Considerations

This article reviews published literature and official guidance and does not report research involving recruited participants or access to identifiable student records. The prospective evaluation protocol describes future research and is not an implemented or ethically approved study.

## Declaration on Generative AI Use

Generative AI assisted literature discovery, source comparison, synthesis, drafting, and revision. Source access and verification limits are reported in the Methodology and Table 1. The revision used retrieved publisher, repository, and official records for targeted checks; this does not constitute independent duplicate human screening or extraction. No synthetic learner observations, new experiments, or unperformed quality assessments are presented as empirical results. The named author retains responsibility for the final manuscript, its interpretations, and compliance with the journal’s disclosure policy.

**Table 1**

*Source inventory and primary interpretive roles*

| Source | Primary role | Access and inference boundary |
|---|---|---|
| Adam et al. (2026) | Game empirical | Publisher abstract; behavior and self-regulation; no causal closure of achievement gaps. |
| Autio et al. (2024) | Official guidance | Official report; risk governance rather than educational effectiveness. |
| Aydemir and Tüzün (2026) | Game empirical | Full text; iterative design evidence, not a definitive efficacy trial. |
| Bastani et al. (2025) | Adjacent empirical | Full text; nongame mathematics tutoring and unassisted performance. |
| Bauer et al. (2025) | Adjacent empirical | Full text; nongame statistics feedback; no significant performance advantage. |
| Cai et al. (2022) | Research synthesis | Publisher abstract; scaffolding effect, not an isolated AI effect. |
| CAST (2024) | Official guidance | Official guidelines; accessibility design, not game-effect evidence. |
| Chen and Chang (2024) | Game empirical | Publisher abstract; randomized science-game comparison with distinct support conditions. |
| Clark et al. (2016) | Research synthesis | Publisher abstract and scholarly record; game design rather than a contemporary AI effect. |
| Corbett and Anderson (1994) | Foundational | Publisher abstract; procedural learner modeling, not game validation. |
| Demedts et al. (2024) | Game empirical | Publisher abstract and repository metadata; cognitive and affective outcomes differ. |
| Demedts et al. (2025) | Game empirical | Publisher abstract; item correctness versus misconception persistence. |
| Faber et al. (2024) | Game empirical | Full text; randomized yoked comparator; principal and exploratory results separated. |
| Hattie and Timperley (2007) | Foundational | Publisher record and accessible scholarly text; feedback framework. |
| Hong et al. (2024) | Research synthesis | Publisher abstract; tailored gamification is not equivalent to an adaptive game. |

**Table 1 (continued)**

*Source inventory and primary interpretive roles*

| Source | Primary role | Access and inference boundary |
|---|---|---|
| Kestin et al. (2025) | Adjacent empirical | Publisher abstract and institutional record; nongame undergraduate physics tutor. |
| Liu and Hwang (2024) | Game empirical | Publisher abstract; structured metalinguistic feedback, not generative-AI validation. |
| Miao and Holmes (2023) | Official guidance | Official UNESCO guidance; human-centered governance, not effectiveness evidence. |
| Ryan and Deci (2020) | Foundational | Full text; motivational theory, not an AI-specific causal estimate. |
| Sailer and Homner (2020) | Research synthesis | Full text; gamification outcomes, not an incremental AI effect. |
| Shen et al. (2024) | Technical survey | Publisher abstract and university repository; knowledge-tracing methods. |
| Shute (2008) | Foundational | Publisher record and visible article extracts; formative-feedback synthesis. |
| UNICEF (2025) | Official guidance | Official version 3.0; child rights, safety, and accountability. |
| Wang et al. (2025) | Game empirical | Publisher abstract; quasi-experiment with fifth-grade science learners. |
| Gupta et al. (2024) | Game empirical | Publisher abstract, revision check; predictive modeling and fairness, not a causal adaptation trial. |
| Huang and Chen (2025) | Game empirical | Publisher abstract, revision check; factorial support comparison; independent delayed mastery not established. |
| Kaşarcı and Yurt (2026) | Research synthesis | Full text via PMC, revision update; broad field context; constituent studies not all independently reviewed here. |
| Panjaburee et al. (2025) | Game empirical | Institutional abstract and metadata, revision check; repeated-measures PLS-SEM; causal benefit not established. |

Note. The inventory contains 28 publications or documents, not necessarily 28 independent datasets. Access labels preserve documented original access and identify revision checks. Publisher abstract or repository record means that full-text review was unavailable or not established; unavailable details are not inferred. The 55 studies within the cited systematic review are not added to this corpus unless individually incorporated.

## Table 2

*Selected game-specific empirical evidence*

| Study and participants | Design and contrast | Verified finding and limit |
|---|---|---|
| Chen & Chang (2024)<br>202 seventh graders | Randomized; game only versus two ChatGPT support conditions. | Examples condition improved achievement versus game only; separate support ingredients. |
| Demedts et al. (2024)<br>288 fourth graders | Random assignment; two feedback games and control. | Near transfer improved in both games; affective results varied by outcome. |
| Demedts et al. (2025)<br>197 fourth graders | Randomized feedback versions; item-level analysis. | Greater correctness improvement with explanatory adaptation; misconception errors did not decrease. |
| Faber et al. (2024)<br>62 medical students | Randomized matched-pairs yoked control. | No detected principal-group advantage; control support often fitted learner needs. |
| Liu & Hwang (2024)<br>76 university students | Comparative grammar-game experiment. | Metalinguistic feedback supported grammar learning; no generative mechanism established. |
| Wang et al. (2025)<br>Fifth graders | Quasi-experimental; LLM-adaptive versus conventional contextual game. | Reported achievement and flow advantages; not randomized; sample count not verified here. |
| Aydemir & Tüzün (2026)<br>70 across five cycles | Design-based research; adaptive programming RPG. | No significant short-term motivation or attitude change; exploratory behavior evidence. |
| Adam et al. (2026)<br>48 junior high students | Quasi-experimental; prior-achievement groups. | Different help-seeking patterns; no demonstrated causal elimination of group differences. |
| Huang & Chen (2025)<br>160 seventh graders | 2 × 2 experimental design; video and chatbot components. | Video-related learning advantage; combined support affected extraneous load. Abstract-level verification; delayed mastery unestablished. |
| Panjaburee et al. (2025)<br>372 lower secondary students | Longitudinal repeated measures; fuzzy logic, decision trees; PLS-SEM. | Motivational pathways and perceived sustainability; not a randomized causal estimate of independent mastery. |
| Gupta et al. (2024)<br>119 middle school learners | Predictive-model evaluation using gameplay and reflection. | Examined prediction and encoded bias; not an instructional-effectiveness experiment. |

Note. Counts are per publication, not unique learners across the corpus. Independence between related Demedts datasets remains unresolved. Randomized, factorial, quasi-experimental, longitudinal associational, design-based, and predictive-model evidence are distinguished. A nonsignificant result does not establish equivalence. LLM = large language model; PLS-SEM = partial least squares structural equation modeling; RPG = role-playing game.

**Table 3**

*Proposed teacher-supervised adaptation cycle*

| Stage | Instructional decision | Required record or control |
|---|---|---|
| Define objective | Specify knowledge or strategy and acceptable independent performance. | Objective, prerequisites, rubric, allowed support. |
| Observe | Separate task response from inferred difficulty. | Task, input, timing context, assistance received. |
| Estimate | Maintain alternative explanations and uncertainty. | Provisional state, confidence and calibration where applicable, competing interpretations, correction and appeal records. |
| Adapt activities | Select relevant practice without permanent low-expectation tracking. | Task dimension changed, rationale, return condition. |
| Adapt pace | Adjust time, review, and progression without equating speed with mastery. | Time allowance, spacing, frequency, duration, advancement rule, access delays, return condition and independent checkpoint. |
| Adapt feedback | Provide bounded support and reduce it when appropriate. | Support level, approved source, model/prompt/RAG/content versions, validation and severity flags. |
| Verify and review | Check independent learning and escalate unresolved cases. | New independent assessment, teacher escalation and override, learner contest, revision history. |

Note. Unvalidated author-developed conceptual framework. These stages are proposed design requirements, not observed outcomes, a validated scoring instrument, or institutional certification.

## Table 4

*Proposed outcomes and operational definitions*

| Measure | Definition or calculation | Interpretation requirement |
|---|---|---|
| Independent mastery | IM = N meeting the prespecified standard without help / N independently assessed. | Report missingness and the prespecified intention-to-treat analysis separately. |
| Misconception persistence | MP = N scored opportunities showing the target misconception / N eligible diagnostic opportunities. | Use independent scoring and new items; report the unit of analysis. |
| Feedback error | FE = N reviewed feedback episodes with substantive error / N reviewed episodes. | Separate factual error from pedagogical mismatch; report severity classes, sampling and uncertainty. |
| Over-assistance | OA = N reviewed episodes exceeding permitted support / N reviewed episodes. | Prespecify allowed support; record help exposure, answer leakage, severity and uncertainty. |
| Retention and transfer | Between-condition score differences on delayed and unfamiliar tasks. | Separate delayed retention, near and far transfer; state timing, novelty, help availability and missingness. |
| Net teacher effort | Preparation + content review + monitoring + dashboard management + verification + correction + escalation time. | Compare equivalent instructional workflows and objectives. |
| Access and equity | Completion, failure, exposure, and outcome distributions by relevant group. | Record device and connectivity limits, accessibility use, hint and advanced-content exposure, and subgroup uncertainty. |
| Adaptation fidelity | AF = N eligible decisions delivered according to the registered policy / N eligible decisions. | Record overrides, technical failures, and comparator overlap. |
| Assisted mastery | AM = N meeting the standard with permitted assistance / N assessed under that support condition. | State units, allowed support, task comparability, and missingness; AM − IM is descriptive unless the design supports causal inference. |

Note. Proposed metrics and operational definitions, not validated instruments or observed results. N denotes a count; define the unit and report every numerator and denominator. Undefined ratios are not estimable. Retain missingness, assistance conditions, and uncertainty. Serious instructional, privacy, or discriminatory errors cannot be compensated by a favorable average score.

## References

Adam, A. S., Hwang, G.-J., & Mubarok, H. (2026). Differentiated impacts of GenAI-supported self-regulated game-based science learning: A behavioural and epistemic network analysis of high- and low-achieving students. Educational Technology Research and Development. Advance online publication. https://doi.org/10.1007/s11423-026-10709-9

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Aydemir, H., & Tüzün, H. (2026). Designing, developing, and evaluating educational games that provide adaptive feedback with generative artificial intelligence. *Turkish Journal of Education, 15*(2), 86–114. https://doi.org/10.19128/turje.1847304

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Bauer, E., Richters, C., Pickal, A. J., Klippert, M., Sailer, M., & Stadler, M. (2025). Effects of AI-generated adaptive feedback on statistical skills and interest in statistics: A field experiment in higher education. *British Journal of Educational Technology, 56*(5), 1735–1757. https://doi.org/10.1111/bjet.13609

Cai, Z., Mao, P., Wang, D., He, J., Chen, X., & Fan, X. (2022). Effects of scaffolding in digital game-based learning on student's achievement: A three-level meta-analysis. *Educational Psychology Review, 34*, 537–574. https://doi.org/10.1007/s10648-021-09655-0

CAST. (2024). CAST Universal Design for Learning Guidelines version 3.0. https://udlguidelines.cast.org/

Chen, C.-H., & Chang, C.-L. (2024). Effectiveness of AI-assisted game-based learning on science learning outcomes, intrinsic motivation, cognitive load, and learning behavior. *Education and Information Technologies, 29*, 18621–18642. https://doi.org/10.1007/s10639-024-12553-x

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

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278. https://doi.org/10.1007/BF01099821

Demedts, F., Kiili, K., Ninaus, M., Lindstedt, A., Reynvoet, B., Sasanguie, D., & Depaepe, F. (2024). The effectiveness of explanatory adaptive feedback within a digital educational game to enhance fraction understanding. *Learning and Instruction, 94*, Article 101976. https://doi.org/10.1016/j.learninstruc.2024.101976

Demedts, F., Said-Metwaly, S., Kiili, K., Ninaus, M., Lindstedt, A., Reynvoet, B., Sasanguie, D., & Depaepe, F. (2025). Adaptive feedback in digital educational games: An explanatory item response theory approach. *Journal of Computer Assisted Learning, 41*(5), Article e70104. https://doi.org/10.1111/jcal.70104

Faber, T. J. E., Dankbaar, M. E. W., van den Broek, W. W., Bruinink, L. J., Hogeveen, M., & van Merriënboer, J. J. G. (2024). Effects of adaptive scaffolding on performance, cognitive load and engagement in game-based learning: A randomized controlled trial. *BMC Medical Education, 24*, Article 943. https://doi.org/10.1186/s12909-024-05698-3

Gupta, A., Carpenter, D., Min, W., Rowe, J., Azevedo, R., & Lester, J. (2024). Detecting and mitigating encoded bias in deep learning-based stealth assessment models for reflection-enriched game-based learning environments. *International Journal of Artificial Intelligence in Education, 34*, 1138–1165. https://doi.org/10.1007/s40593-023-00379-6

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hong, Y., Saab, N., & Admiraal, W. (2024). Approaches and game elements used to tailor digital gamification for learning: A systematic literature review. *Computers & Education, 212*, Article 105000. https://doi.org/10.1016/j.compedu.2024.105000

Huang, K., & Chen, C.-H. (2025). Instructional video and GenAI-supported chatbot in digital game-based learning: Influences on science learning, cognitive load and game behaviours. *Journal of Computer Assisted Learning, 41*(4), Article e70094. https://doi.org/10.1111/jcal.70094

Kaşarcı, İ., & Yurt, E. (2026). Effectiveness of AI-supported game-based learning: A systematic review of outcomes, challenges, and future directions. *Behavioral Sciences, 16*(7), Article 1050. https://doi.org/10.3390/bs16071050

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*, Article 17458. https://doi.org/10.1038/s41598-025-97652-6

Liu, L.-A., & Hwang, G.-J. (2024). Effects of metalinguistic corrective feedback on novice EFL students' digital game-based grammar learning performances, perceptions and behavioural patterns. *British Journal of Educational Technology, 55*(2), 687–711. https://doi.org/10.1111/bjet.13400

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Panjaburee, P., Hwang, G.-J., Intarakamhang, U., & Srisawasdi, N. (2025). Modeling the sustainability perspectives on personalized digital games for digital citizenship education: A PLS-SEM approach. *Computers and Education: Artificial Intelligence, 9*, Article 100498. https://doi.org/10.1016/j.caeai.2025.100498

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology, 61*, Article 101860. https://doi.org/10.1016/j.cedpsych.2020.101860

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

Shen, S., Liu, Q., Huang, Z., Zheng, Y., Yin, M., Wang, M., & Chen, E. (2024). A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies, 17*, 1898–1919. https://doi.org/10.1109/TLT.2024.3383325

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

UNICEF. (2025). Guidance on AI and children (Version 3.0). https://www.unicef.org/innocenti/reports/policy-guidance-ai-children

Wang, M., Zhang, D., Zhu, J., & Gu, H. (2025). Effects of incorporating a large language model-based adaptive mechanism into contextual games on students' academic performance, flow experience, cognitive load and behavioral patterns. *Journal of Educational Computing Research, 63*(3), 662–694. https://doi.org/10.1177/07356331251321719