PRACTICE PACKAGE DESIGN IN COMPUTER-ASSISTED LEARNING: INTRINSIC INTEREST AND DELAYED RETENTION

Vasileios Daloukas[1], Anthi Karatrantou[1], George Koutromanos[2] and Christos Panagiotakopoulos[1]

[1]Department of Educational Sciences and Social Work, University of Patras, Greece
[2]Department of Pedagogy and Primary Education, National and Kapodistrian University of Athens, Greece

ABSTRACT

Digital practice packages combine feedback, reattempts, and performance cues, yet little is known about how post-activity interest relates to later retention. Guided by self-determination theory, we compared a digital quiz package with a paper worksheet using the same 36-item pool in 30 fifth-grade classrooms from 24 primary schools (N = 437). Within classrooms, two algorithmically balanced groups were formed and randomly assigned to conditions. Students completed a pretest, practice activity, intrinsic-interest measure, and delayed posttest 7–10 days later. In the available-outcome analysis (N = 397), students assigned to the digital package showed higher delayed retention after adjustment for pretest; the classroom-clustered estimate was B = 2.62, 95% confidence interval (CI) [1.75, 3.50]. In complete-case mediation (N = 391), the digital package was associated with higher intrinsic interest (B = 0.95, 95% CI [0.74, 1.17]), and the estimated indirect effect was positive (B = 1.21, 95% CI [0.40, 2.09]). Because the mediator was not randomised, the pathway is associational, not causal.

KEYWORDS

Computer-assisted learning, intrinsic interest, delayed retention, digital practice, mediation analysis

1. INTRODUCTION

Durable learning from practice depends not only on exposure to content but also on what learners are required to do with it. Retrieval-practice research indicates that actively retrieving previously studied material can strengthen subsequent retention across a wide range of authentic educational settings [1–3]. Classroom experiments have likewise demonstrated long-term benefits of low-stakes testing with feedback in sixth-grade social studies [4] and U.S. History [5]. Particularly close to the present context, Franzoi et al. [6] recently found that retrieval practice with accuracy feedback and repeated correction improved fifth-grade pupils' learning of authentic History materials relative to rereading in a real primary-school setting. This evidence provides a rationale for examining practice packages that combine retrieval, corrective feedback, and opportunities to reattempt errors.

Corrective feedback is commonly incorporated into retrieval-based classroom practice [1–6], and repeated correction of errors has also been used successfully with primary-school pupils [6]. In a digital setting, feedback can be delivered immediately and paired with reattempts, optional aids, points, rankings, and competitive displays. This creates a multi-component practice environment, but it also means that several instructional features vary at the same time. Meta-analytic work

generally reports positive effects of educational gamification, together with substantial between-study variation [7, 8]. Li et al. [9] found the clearest motivational benefits for intrinsic motivation, autonomy, and relatedness, with much weaker evidence for perceived competence. A meta-analysis of Kahoot!-based interventions similarly reported positive average effects on achievement, retention, and motivation, again with marked heterogeneity [10]. We therefore examined delayed learning and students' motivational experience under the two practice packages.

1.1. Post-Activity Interest/Enjoyment as a Candidate Associational Pathway

Students may practise identical material but experience the activity differently, so post-activity motivational states are plausible correlates of later learning. Situational interest is typically understood as an activity- or context-elicited state that can support attention, persistence, and deeper engagement [11, 12]. Ainley, Hidi, and Berndorff [13] linked interest with affective reactions and persistence, which were in turn associated with learning. Rotgans and Schmidt [14] similarly found that situational interest was related to achievement-relevant classroom behaviour and academic performance.

Relatedly, within self-determination theory (SDT), intrinsic motivation refers to engaging in an activity because the activity itself is experienced as interesting or enjoyable, rather than primarily to obtain external rewards or avoid negative consequences [15, 16]. In classroom terms, an intrinsically motivated pupil participates because the learning task is satisfying in its own right, not merely to earn points, attain a rank, or comply with an instruction. This form of motivation may support attention and persistence, although enjoyment alone does not ensure learning. The Interest/Enjoyment subscale of the Intrinsic Motivation Inventory (IMI) is a commonly used self-report indicator of intrinsic motivation [17, 18]. In the present study, it represents pupils' immediate motivational experience after practice, not a stable disposition or a causally identified mechanism.

Evidence from digital learning is mixed rather than uniformly supportive of a motivational pathway. In elementary-school science, Rayan and Watted [19] observed gains in conceptual learning alongside higher interest and enjoyment during Kahoot!-supported instruction. Cai et al. [20] also reported an indirect statistical pathway through intrinsic motivation in a digital game-based learning study, although their participants were university students and their experimental contrast concerned collaboration rather than the practice features studied here. Other findings show that motivational and learning outcomes need not move together. Kuklick [21], for example, found no overall learning or affective-motivational benefit from giving learners control over whether immediate automated feedback was accessed, although learners who did not retrieve the feedback reported a less positive emotional state. The mixed findings warrant examining Interest/Enjoyment as a candidate pathway, without presuming that higher enjoyment is accompanied by better learning.

SDT also suggests that the digital package need not affect motivation in a single direction. Immediate informational feedback and optional aids could support competence or autonomy, while rankings, points, and competitive cues could be experienced as evaluative or controlling, depending on the learner. The Interest/Enjoyment score therefore captures one aspect of students' motivational response to the package as a whole [15, 16].

1.2. The Present Study

Retrieval-practice studies have established benefits for retention in classroom settings [1–6], whereas educational gamification research often examines motivation alongside achievement [7–10]. Less clear is what occurs when a digital practice package combining retrieval, corrective feedback, reattempts, and game-like performance cues is compared with a paper activity using the same 36-item pool in primary-school classrooms and retention is assessed after a delay. Franzoi et al. [6] provide a particularly close age- and subject-relevant example of retrieval practice with corrective feedback, but their study did not test a gamified digital package or examine an indirect pathway through post-activity motivation.

Accordingly, we compared two practice packages using the same 36-item pool in fifth-grade History classrooms. The digital condition combined immediate corrective feedback, error-contingent reattempts, optional aids, and performance cues, whereas the paper worksheet was completed in a single pass and feedback was provided after completion. Because these delivery features changed together, the randomised contrast concerns the package as implemented and cannot identify the effect of any one component. We also examined post-activity Interest/Enjoyment because it is the IMI indicator most directly aligned with intrinsic motivation.

1.3. Research Hypotheses

We tested two hypotheses:

H1. The classroom-implemented digital practice package was expected to yield higher delayed retention than the paper worksheet using the same 36-item pool.
H2. The digital practice package was expected to be associated with higher post-activity intrinsic interest, and the pattern of estimates was expected to be consistent with a positive indirect pathway from practice-package assignment to delayed retention through intrinsic interest.

2. MATERIALS AND METHODS

2.1. Participants

Our sample comprised 437 fifth-grade students (approximately 11 years old) from 30 classrooms in 24 public primary schools in Western Greece. Written parental consent was obtained, and participation was voluntary and confidential. The research team received only six-digit participant codes; teachers retained the code–student key locally. The study was approved by the Administrative Board of the Department of Educational Sciences and Social Work, University of Patras (meeting no. 2/31 January 2023; protocol no. 13407, 21 February 2023). No a priori power analysis was conducted; sample size was determined pragmatically by classroom availability.

For H1, students were required to provide at least 33 of 36 valid responses at both pretest and delayed posttest, yielding N = 397; IMI availability was not required. The mediation sample additionally required at least six of the seven Interest/Enjoyment items, yielding N = 391. All 30 classrooms remained represented. Package-specific inclusion rates and differential-inclusion analyses are reported in the Appendix (Table 4).

2.2. Design and Procedure

Students completed the curriculum-aligned pretest during the first class period, administered by the researcher with the classroom teacher present. During the second 45-minute period, the digital group moved to the computer laboratory and the paper group remained in the classroom;

approximately 30 minutes were available for practice. The groups then reconvened and completed the IMI Interest/Enjoyment measure immediately after practice. The classroom teacher administered the delayed posttest 7–10 days later following standardised administration instructions.

Within each classroom, an exhaustive-search deterministic algorithm first formed two near-equal groups by minimising imbalance in group size, gender composition, and the means and standard deviations of teacher-estimated History achievement, both overall and by gender. The algorithm formed the groups but did not assign conditions. After group membership was fixed, a separate random draw within each classroom assigned one whole group to the digital quiz package and the other to the paper worksheet; no student was subsequently reassigned. Pretest scores were unavailable at allocation and were not used. Among students with an eligible pretest (N = 428), the baseline package difference was −0.07 correct items (classroom-cluster bootstrap 95% CI [−1.15, 0.95]; standardised mean difference = −0.01). Gender composition was also similar in the full randomised sample (54.4% boys in the digital package and 52.3% in the paper package; standardised mean difference = 0.04).

2.3. Learning Environment

Students worked individually, without verbal collaboration or content-related assistance from the researcher or teacher. The digital package combined immediate corrective feedback, error-contingent reattempts, optional aids, points, rank-position cues, and an asynchronous duel display; these performance cues correspond to commonly studied gamification elements [8, 22, 23]. The paper worksheet contained the same 36 items, was completed in a single pass, and was self-scored from an answer sheet after completion. Table 1 summarises the implemented contrast.

Table 1. Components of the implemented practice packages (digital quiz package versus paper worksheet)

Package component	**Digital quiz package**	**Paper worksheet**
Feedback	Immediate corrective feedback after each response	Answer sheet after completion
Attempts	Incorrect items were re-presented	Single pass; no reattempts
Performance cues	+3 for a correct response, −1 for an incorrect response (minimum 0), +3 duel bonus when the opponent was incorrect, rank/leaderboard, and asynchronous duel display	Final self-score only; no points, ranking, or competitive display during practice
Optional aids	Available on demand during practice	Not available
Format	Individual computer-based interface	Individual paper worksheet

Note. Both packages used the same curricular topic and 36-item pool. Digital practice was supervised by the researcher in the computer laboratory, whereas paper practice was supervised by the classroom teacher. Both conditions followed standardised instructions.

2.4. Technology and Implementation

To preserve privacy and standardise delivery and scoring, a code-based system produced individualised printed versions of the pretest, paper worksheet, and delayed posttest. Item and option orders were deterministically permuted using the classroom code, six-digit participant code, and form type. The research team used predefined scripts to reconstruct the mapping and apply the same prespecified answer key across forms.

2.5. Measures

2.5.1. Learning Performance

Prior knowledge and delayed retention were assessed with the same 36-item, four-option multiple-choice set aligned with the national History curriculum; using the same set was intended to index retention of the practised items. Scores were the number of correct responses (range 0–36). Internal consistency, estimated using Kuder–Richardson Formula 20 (KR-20) [24], was acceptable at pretest (KR-20 = 0.77) and high at delayed posttest (KR-20 = 0.92) in the complete-case mediation sample (N = 391). Items were authored by the research team to align with the standard textbook and piloted with fifth-grade students who were not included in the main sample.

2.5.2. IMI Interest/Enjoyment

Post-activity intrinsic interest was measured immediately after practice with the seven-item IMI Interest/Enjoyment subscale [17, 18]. Responses used a 5-point scale (1 = strongly disagree, 5 = strongly agree), and the score was the item mean when at least six items were answered. The items were translated into Greek and independently back-translated following cross-cultural adaptation principles [25]. Internal consistency was high (Cronbach's α = 0.92) among students who answered all seven items (N = 366).

2.5.3. Data Processing and Analysis Plan

Analyses were conducted in R version 4.6.1 [26]. Practice-package assignment was coded 0 = paper worksheet and 1 = digital quiz package; descriptive statistics were unadjusted. H1 was tested in the available-outcome sample (N = 397) using ordinary least squares (OLS) regression of delayed posttest score on assignment and pretest score. Inference used classroom-clustered standard errors (SEs) with a bias-reduced linearisation adjustment (CR2) and Satterthwaite degrees of freedom [27–29]. Robustness was assessed using classroom fixed effects, school-level clustering, design-consistent randomisation inference, and full-sample multiple imputation.

The secondary mediation analysis used the complete-case sample (N = 391) and two OLS models: Interest/Enjoyment as a function of assignment and pretest score, and delayed posttest score as a function of assignment, Interest/Enjoyment, and pretest score. The indirect estimate was the product of the a- and b-path coefficients ($a \times b$). Individual paths used CR2 inference; percentile confidence intervals for the indirect and same-sample total effects used 5,000 classroom-cluster bootstrap resamples. Because Interest/Enjoyment was measured after practice and was not randomised, the indirect estimate is interpreted as an associational pathway rather than a causally identified effect. A residual-correlation sensitivity analysis assessed departures from sequential ignorability [30, 31].

Missing-data sensitivity analyses used multiple imputation by chained equations with predictive mean matching (50 datasets; 30 iterations) and stabilised inverse-probability weighting (IPW). The Appendix reports specifications, diagnostics, and additional checks involving device ownership, classroom fixed effects, leave-one-classroom-out stability, design-consistent randomisation inference, and school-level dependence. The mediation model was theory-guided but not preregistered; the package × Interest/Enjoyment interaction was a post hoc specification check.

3. RESULTS

3.1. Descriptive Statistics and Correlations

Table 2 presents descriptive statistics and bivariate correlations among the core variables.

Table 2. Descriptive statistics and correlations among core variables (N = 391)

Variable	*M*	*SD*	Range	1	2	3
1. Pretest score	18.85	5.45	5–36	—		
2. Delayed posttest score	22.84	8.43	2–36	.70	—	
3. IMI Interest/Enjoyment	3.65	1.00	1–5	.20	.33	—

Note. *M* = mean; *SD* = standard deviation. Pearson correlations are reported below the diagonal; dashes indicate self-correlations.

Figure 1 summarises the focal mediation structure, and Table 3 reports the component paths together with the implied indirect effect, adjusted package coefficient, and total effect. The Appendix provides further details on sample construction, missing-data sensitivity, and robustness checks.

Response completeness was very high across the two assessments (mean non-omitted responses = 35.85/36) and was similar across practice packages. The package difference in non-omitted responses (out of 36 items) was 0.02 at pretest (95% CI [−0.09, 0.13], p = .709) and 0.04 at delayed posttest (95% CI [−0.05, 0.13], p = .380), using CR2 classroom-clustered standard errors.

3.2. Primary Package Effect on Delayed Retention (H1)

In the primary available-outcome analysis (N = 397), students assigned to the digital quiz package scored higher on the delayed posttest than students assigned to the paper worksheet, after adjustment for pretest score (B = 2.62, SE = 0.42, 95% CI [1.75, 3.50], t(25.50) = 6.18, p < .001). The estimate was essentially unchanged in the classroom fixed-effects specification (B = 2.63, 95% CI [1.67, 3.59]) and when CR2 inference was clustered at the school level (B = 2.62, 95% CI [1.73, 3.52]). Design-consistent randomisation inference based on 5,000 whole-group label swaps within classrooms likewise yielded p < .001. A full-randomised-sample multiple-imputation analysis (N = 437) produced a similar estimate (B = 2.83, SE = 0.59, 95% CI [1.62, 4.05], p < .001).

3.3. Secondary Indirect-Effect Model

In the secondary complete-case mediation analysis (N = 391), practice-package assignment was positively associated with post-activity IMI Interest/Enjoyment (a = 0.95, 95% CI [0.74, 1.17]). In the outcome model, Interest/Enjoyment remained positively associated with delayed posttest performance after adjustment for practice-package assignment and pretest score (b = 1.27, 95% CI [0.36, 2.18]), while the adjusted package coefficient remained positive (c' = 1.51, 95% CI [0.21, 2.81]). The implied indirect effect through Interest/Enjoyment was positive (B = 1.21, 95% CI [0.40, 2.09]). For same-sample decomposition, the total package coefficient in this N = 391 mediation sample was B = 2.72 (95% CI [1.92, 3.46]); this quantity is not the primary H1 estimate. Table 3 reports the full secondary mediation model, and Figure 1 summarises the focal pathway.

Table 3. Core indirect-effect model (IMI Interest/Enjoyment pathway)

Effect	*B*	β	*SE* (CR2)	95% CI	*t*	*df*	*p*
Package → IMI Interest/Enjoyment (*a*)	0.95	0.48	0.11	[0.74, 1.17]	8.99	25.51	< .001
Pretest → IMI	0.04	0.20	0.01	[0.02, 0.05]	5.06	21.23	< .001
IMI Interest/Enjoyment → Delayed posttest (*b*)	1.27	0.15	0.44	[0.36, 2.18]	2.89	22.55	.008
Package → Delayed posttest (*c′*)	1.51	0.09	0.63	[0.21, 2.81]	2.39	23.65	.025
Pretest → Delayed posttest	1.04	0.67	0.06	[0.92, 1.15]	18.35	21.38	< .001
Indirect effect (*a* × *b*)	1.21	—	—	[0.40, 2.09]	—	—	—
Same-sample total effect (*c*; mediation *N*)	2.72	—	—	[1.92, 3.46]	—	—	—

Note. Coefficients are unstandardised B; β is reported for single-equation paths. Standard errors for *a*-, *b*-, *c′*-paths and the pretest covariate are CR2 cluster-robust with Satterthwaite degrees of freedom (classroom as the clustering unit). Practice-package assignment is coded 0 = paper worksheet, 1 = digital quiz package. Pretest is included as a covariate in all equations. For total and indirect effects, we report point estimates and percentile 95% CIs from a classroom-cluster bootstrap with 5,000 resamples; we do not report CR2 SEs or t-tests for these quantities. The total-effect row is retained only for same-sample mediation decomposition (N = 391) and is not the primary H1 estimate; primary H1 uses the available-outcome sample (N = 397). *SE* = standard error; CI = confidence interval; df = degrees of freedom.

3.4. Robustness and Specification Checks

Missing-data sensitivity analyses were broadly directionally consistent with the complete-case mediation results (Appendix). The indirect-effect point estimate remained positive under both multiple imputation (B = 0.90, 95% CI [−0.03, 1.84]) and inverse probability weighting (B = 1.17, 95% CI [0.27, 2.07]). However, the multiple-imputation confidence interval included zero, indicating that evidence for the indirect pathway was less robust to assumptions about missing data than the primary package effect on delayed retention.

As a post hoc specification check, adding a practice-package × IMI Interest/Enjoyment interaction to the delayed-posttest model yielded little evidence of moderation (B = 0.90, SE = 0.68, 95% CI [−0.51, 2.30], $t(20.92) = 1.32$, $p = .200$; $\Delta R^2 = .002$). The corresponding classroom-fixed-effects specification yielded a similar result (B = 0.86, 95% CI [−0.69, 2.42], $p = .262$; $\Delta R^2 = .002$).

Because six of the 24 participating schools contributed two classrooms each, we repeated the focal analyses using school-clustered CR2 inference and a school-cluster bootstrap. Estimates closely matched the classroom-clustered results for both the primary H1 effect and the secondary mediation model (Appendix).

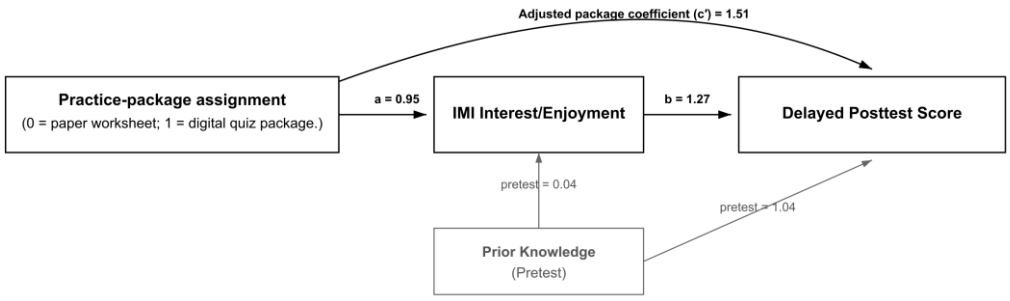

Figure 1. Conceptual mediation model linking practice-package assignment to delayed posttest score via post-activity IMI Interest/Enjoyment (adjusting for pretest); path labels correspond to unstandardised coefficients (B); inference details are provided in Table 3

4. DISCUSSION

The primary finding was an adjusted delayed-posttest advantage of 2.62 items for the digital practice package. This result is consistent with classroom evidence that retrieval-based practice can support longer-term learning [1–6], including recent work with fifth-grade History materials [6]. However, the present intervention bundled retrieval with immediate feedback, reattempts, optional aids, and gamified performance cues. The finding should therefore be interpreted as a package-level contrast rather than evidence that any single component produced the advantage.

In the mediation sample, the package coefficient remained positive after Interest/Enjoyment was entered in the delayed-posttest model. Thus, the motivational measure was associated with only part of the observed package difference. Prior retrieval-practice research, including classroom work incorporating feedback and repeated correction [3, 6], makes these features plausible contributors, but the current design cannot separate their effects from gamified cues, digital delivery, or the other elements that differed between conditions.

The higher Interest/Enjoyment scores in the digital condition are also compatible with previous work showing motivational benefits in some gamified learning environments. Rayan and Watted [19] reported concurrent improvements in learning, interest, and enjoyment among elementary-school pupils using Kahoot!, and Li et al. [9] found positive average effects of gamification on intrinsic motivation. Cai et al. [20] further reported an indirect statistical pathway through intrinsic motivation in a university digital game-based learning study, although the intervention and participant population differed substantially from those examined here. Earlier work has also linked interest with affective engagement, persistence, and learning [13]. Practice-package assignment was randomised; Interest/Enjoyment was not, and the post-activity score may partly reflect perceived success or other treatment-induced experiences. The positive indirect estimate is therefore interpreted as an associational pattern rather than evidence that increased interest causally improved delayed retention.

The direction of the motivational result should not be taken to mean that every element of the digital package was intrinsically motivating. Under SDT, informational feedback and optional aids could support competence or autonomy, whereas rankings, points, and competitive cues could create evaluative pressure. Interest/Enjoyment thus reflects students' response to the package as a whole rather than to any particular feature. A factorial design that varies these elements separately would be needed to determine whether supportive and potentially controlling features make distinct contributions to students' motivational experience.

4.1. Theoretical and Practical Implications

The results are compatible with a motivational pathway involving Interest/Enjoyment, but they do not demonstrate causal mediation. From a practical perspective, the evidence concerns the digital package as a whole. It therefore remains unclear whether immediate feedback, reattempts, optional aids, or performance cues would produce similar effects if implemented separately. Classroom use should also take into account device availability, accessibility requirements, and local attitudes towards competitive features.

4.2. Limitations

The findings are subject to several design limitations. First, because delivery elements co-occurred within each package, the study cannot isolate the contribution of any single feature to the observed effects. Detailed process measures of individual reattempt frequency, optional-aid use, and practice duration were not included in the present analysis.

Second, the mediator was not randomised; thus, causal interpretation of the indirect effect would require the sequential-ignorability assumption. In the Imai et al. residual-correlation sensitivity analysis [30, 31], the estimated indirect effect was reduced to approximately zero at a residual mediator–outcome error correlation of $\rho = .185$ (R^{2*} product = .034; $R^{2\sim}$ product = .012). This indicates that the indirect estimate is sensitive to departures from sequential ignorability. We therefore interpret the mediation results as a theory-consistent associational pathway rather than evidence of a causal mediating mechanism. Moreover, no baseline measure of Interest/Enjoyment was available. Consequently, despite adjustment for pretest achievement and the sensitivity analysis, residual mediator–outcome confounding by stable or pre-existing learner characteristics cannot be ruled out.

Third, situational motivation may also have been influenced by contextual factors unrelated to the intended delivery features, including differences in physical setting and possible research-participation or observer-reactivity effects [32, 33].

Because Interest/Enjoyment was measured after practice, post-treatment experiences such as perceived success, competence, or immediate learning could have influenced both the mediator and delayed performance. Treatment-induced mediator–outcome confounding therefore cannot be excluded.

Physical setting and supervisor also differed between conditions, so their effects cannot be separated from the overall package contrast.

In addition, because the quiz package was computer-delivered whereas the worksheet was paper-based, part of the Interest/Enjoyment difference may reflect novelty or a medium-related response. The analyses allowed for within-classroom dependence through clustered inference and design-consistent randomisation inference; however, these procedures do not rule out possible interference between students exposed to the duel/rank layer.

Finally, the delayed posttest intentionally sampled the practised items to index retention of the practised set; this outcome does not speak to transfer beyond the practised pool. Because the same item set was used at pretest and delayed posttest, repeated exposure may also have contributed to retention of the tested material. This exposure was common to both conditions and therefore does not by itself explain the between-package difference, but it further limits generalisation beyond the practised item set. Transfer beyond the practised material is less certain: retrieval practice shows a positive average transfer effect, but its magnitude varies substantially across transfer types and study conditions [34].

5. CONCLUSIONS AND FUTURE RESEARCH

In these fifth-grade classrooms, assignment to the digital practice package was associated with an adjusted advantage of 2.62 correct items on the 36-item delayed posttest relative to the paper worksheet using the same 36-item pool, and this package-level difference was robust across the primary sensitivity analyses. Students in the digital condition also reported greater post-activity Interest/Enjoyment, and the complete-case analysis yielded a positive indirect estimate through this measure. However, the indirect pattern was less robust under multiple-imputation

assumptions and should not be interpreted causally because Interest/Enjoyment was not randomised. The findings therefore support the digital package as implemented while leaving unresolved the roles of motivation and individual package components.

Future studies could separate the components of the digital package by varying immediate feedback, reattempts, optional aids, and competitive cues while keeping the delivery medium constant. Including a pre-activity measure of motivation would help distinguish pre-existing differences from responses to the intervention, while process logs could document item exposure, reattempts, aid use, and time on task. It would also be useful to examine transfer to unpractised material and to replicate the study across subjects, age groups, schools, and different levels of device access. Experimental manipulation of the features expected to influence motivation would provide a stronger test of whether motivational change contributes to delayed retention.

DATA AVAILABILITY STATEMENT

De-identified data, analysis scripts, scoring and mapping source code, and generated analysis outputs are available on Figshare at https://doi.org/10.6084/m9.figshare.33297534

CONFLICTS OF INTEREST

Christos Panagiotakopoulos serves as Editor-in-Chief of the *International Journal on Integrating Technology in Education* (IJITE), and Anthi Karatrantou serves as a member of the journal's Editorial Board. These editorial roles are disclosed as potential conflicts of interest in relation to this manuscript. The other authors declare no conflicts of interest.

FUNDING

This research received no external funding.

REFERENCES

[1] P. K. Agarwal, L. D. Nunes, and J. R. Blunt, "Retrieval Practice Consistently Benefits Student Learning: a Systematic Review of Applied Research in Schools and Classrooms", Educ Psychol Rev, vol. 33, no. 4, pp. 1409–1453, Dec. 2021, doi: 10.1007/s10648-021-09595-9.

[2] O. O. Adesope, D. A. Trevisan, and N. Sundararajan, "Rethinking the Use of Tests: A Meta-Analysis of Practice Testing", Review of Educational Research, vol. 87, no. 3, pp. 659–701, June 2017, doi: 10.3102/0034654316689306.

[3] C. Yang, L. Luo, M. A. Vadillo, R. Yu, and D. R. Shanks, "Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review", Psychological Bulletin, vol. 147, no. 4, pp. 399–435, Apr. 2021, doi: 10.1037/bul0000309.

[4] H. L. Roediger, P. K. Agarwal, M. A. McDaniel, and K. B. McDermott, "Test-enhanced learning in the classroom: Long-term improvements from quizzing", Journal of Experimental Psychology: Applied, vol. 17, no. 4, pp. 382–395, 2011, doi: 10.1037/a0026252.

[5] S. K. Carpenter, H. Pashler, and N. J. Cepeda, "Using tests to enhance 8th grade students' retention of U.S. history facts", Applied Cognitive Psychology, vol. 23, no. 6, pp. 760–771, Sept. 2009, doi: 10.1002/acp.1507.

[6] L. Franzoi, V. Cembrani, C. Mulatti, and B. Treccani, "Retrieval practice enhances learning in real primary school settings, whether distributed or not", Front. Psychol., vol. 16, p. 1632206, Aug. 2025, doi: 10.3389/fpsyg.2025.1632206.

[7] S. Bai, K. F. Hew, and B. Huang, "Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts", Educational Research Review, vol. 30, p. 100322, June 2020, doi: 10.1016/j.edurev.2020.100322.

[8] M. Sailer and L. Homner, "The Gamification of Learning: a Meta-analysis", Educ Psychol Rev, vol. 32, no. 1, pp. 77–112, Mar. 2020, doi: 10.1007/s10648-019-09498-w.
[9] L. Li, K. F. Hew, and J. Du, "Gamification enhances student intrinsic motivation, perceptions of autonomy and relatedness, but minimal impact on competency: a meta-analysis and systematic review", Educational Technology Research and Development, vol. 72, no. 2, pp. 765–796, Apr. 2024, doi: 10.1007/s11423-023-10337-7.
[10] O. Özdemir, "Kahoot! Game-based digital learning platform: A comprehensive meta-analysis", Journal of Computer Assisted Learning, vol. 41, no. 1, p. e13084, Feb. 2025, doi: 10.1111/jcal.13084.
[11] M. Ainley, "Students' Interest and Engagement in Classroom Activities", in Handbook of Research on Student Engagement, S. L. Christenson, A. L. Reschly, and C. Wylie, Eds, Boston, MA: Springer US, 2012, pp. 283–302. doi: 10.1007/978-1-4614-2018-7_13.
[12] S. Hidi and K. A. Renninger, "The Four-Phase Model of Interest Development", Educational Psychologist, vol. 41, no. 2, pp. 111–127, June 2006, doi: 10.1207/s15326985ep4102_4.
[13] M. Ainley, S. Hidi, and D. Berndorff, "Interest, learning, and the psychological processes that mediate their relationship", Journal of Educational Psychology, vol. 94, no. 3, pp. 545–561, Sept. 2002, doi: 10.1037/0022-0663.94.3.545.
[14] J. I. Rotgans and H. G. Schmidt, "Situational interest and academic achievement in the active-learning classroom", Learning and Instruction, vol. 21, no. 1, pp. 58–67, Feb. 2011, doi: 10.1016/j.learninstruc.2009.11.001.
[15] R. M. Ryan and E. L. Deci, Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. Guilford Press, 2017. doi: 10.1521/978.14625/28806.
[16] E. L. Deci and R. M. Ryan, "The 'What' and 'Why' of Goal Pursuits: Human Needs and the Self-Determination of Behavior", Psychological Inquiry, vol. 11, no. 4, pp. 227–268, Oct. 2000, doi: 10.1207/S15327965PLI1104_01.
[17] E. McAuley, T. Duncan, and V. V. Tammen, "Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis", Research Quarterly for Exercise and Sport, vol. 60, no. 1, pp. 48–58, Mar. 1989, doi: 10.1080/02701367.1989.10607413.
[18] R. M. Ryan, "Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory", Journal of Personality and Social Psychology, vol. 43, no. 3, pp. 450–461, Sept. 1982, doi: 10.1037/0022-3514.43.3.450.
[19] B. Rayan and A. Watted, "Enhancing Education in Elementary Schools through Gamified Learning: Exploring the Impact of Kahoot! on the Learning Process", Education Sciences, vol. 14, no. 3, p. 277, Mar. 2024, doi: 10.3390/educsci14030277.
[20] Z. Cai et al., "Effect of Collaboration in Digital Game-Based Learning: The Roles of Flow Experience and Intrinsic Motivation", Journal of Computer Assisted Learning, vol. 41, no. 4, p. e70065, Aug. 2025, doi: 10.1111/jcal.70065.
[21] L. Kuklick, "Effects of learner choice over automated, immediate feedback", Learning and Instruction, vol. 96, p. 102065, Apr. 2025, doi: 10.1016/j.learninstruc.2024.102065.
[22] Y. Hong, N. Saab, and W. Admiraal, "Approaches and game elements used to tailor digital gamification for learning: A systematic literature review", Computers & Education, vol. 212, p. 105000, Apr. 2024, doi: 10.1016/j.compedu.2024.105000.
[23] J. Wang, S. Gong, Y. Cao, X. Guo, and P. Peng, "Personalization in educational gamification: Learners with different trait competitiveness benefit differently from rankings on leaderboards", Computers & Education, vol. 225, p. 105196, Feb. 2025, doi: 10.1016/j.compedu.2024.105196.
[24] G. F. Kuder and M. W. Richardson, "The Theory of the Estimation of Test Reliability", Psychometrika, vol. 2, no. 3, pp. 151–160, Sept. 1937, doi: 10.1007/BF02288391.
[25] D. E. Beaton, C. Bombardier, F. Guillemin, and M. B. Ferraz, "Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures", Spine, vol. 25, no. 24, pp. 3186–3191, Dec. 2000, doi: 10.1097/00007632-200012150-00014.
[26] R Core Team, R: A Language and Environment for Statistical Computing. (2026). R Foundation for Statistical Computing, Vienna, Austria. doi: 10.32614/R.manuals.
[27] R. M. Bell and D. McCaffrey, "Bias reduction in standard errors for linear regression with multi-stage samples", Survey Methodology, vol. 28, no. 2, pp. 169–181, 2002.
[28] J. E. Pustejovsky and E. Tipton, "Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models", Journal of Business & Economic Statistics, vol. 36, no. 4, pp. 672–683, Oct. 2018, doi: 10.1080/07350015.2016.1247004.

[29] J. E. Pustejovsky, clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. (2026). R. CRAN. doi: 10.32614/CRAN.package.clubSandwich.
[30] K. Imai, L. Keele, and D. Tingley, "A general approach to causal mediation analysis", Psychological Methods, vol. 15, no. 4, pp. 309–334, 2010, doi: 10.1037/a0020761.
[31] K. Imai, L. Keele, and T. Yamamoto, "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects", Statist. Sci., vol. 25, no. 1, pp. 51–71, Feb. 2010, doi: 10.1214/10-STS321.
[32] J. McCambridge, J. Witton, and D. R. Elbourne, "Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects", Journal of Clinical Epidemiology, vol. 67, no. 3, pp. 267–277, Mar. 2014, doi: 10.1016/j.jclinepi.2013.08.015.
[33] W. R. Shadish, T. D. Cook, and D. T. Campbell, Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin Company, 2002.
[34] S. C. Pan and T. C. Rickard, "Transfer of test-enhanced learning: Meta-analytic review and synthesis", Psychological Bulletin, vol. 144, no. 7, pp. 710–756, July 2018, doi: 10.1037/bul0000151.

APPENDIX. ANALYTIC SAMPLE AND ROBUSTNESS CHECKS

Applying the achievement-test criterion (at least 33 of 36 valid responses at both assessments) yielded $N = 397$ for H1. Requiring at least six of seven IMI Interest/Enjoyment items yielded $N = 391$ for mediation. Inclusion was high and similar across packages (Table 4). The odds ratio (*OR*) from a logistic model with classroom-clustered CR2 inference was 1.10 (95% CI [0.60, 2.01], $p = .748$) for H1 inclusion and 0.99 (95% CI [0.18, 5.63], $p = .995$) for IMI completeness conditional on H1 eligibility.

Table 4. Analytic-sample availability by practice package

Analysis sample	**Digital**	**Paper**	**Overall**
H1 available-outcome	198/217 (91.2%)	199/220 (90.5%)	397/437 (90.8%)
Mediation complete-case	195/217 (89.9%)	196/220 (89.1%)	391/437 (89.5%)

For the full-randomised-sample H1 analysis, multiple imputation by chained equations with predictive mean matching used 50 datasets and 30 iterations. The imputation model included assignment, gender, classroom membership as a categorical predictor, assessment and IMI response-completeness measures, pretest score, delayed-posttest score, and Interest/Enjoyment. Observed test scores were retained whenever the corresponding completeness criterion was met. Each completed dataset was analysed using classroom-clustered CR2 inference, and estimates were pooled with the Barnard–Rubin small-sample adjustment. The pooled estimate was $B = 2.83$, $SE = 0.59$, 95% CI [1.62, 4.05], $t(22.26) = 4.84$, $p < .001$.

For mediation, multiple imputation (MI) and stabilised inverse-probability weighting (IPW) were complementary sensitivity analyses (Table 5). IPW used assignment in the numerator model and additionally included baseline pretest score, valid pretest-response count, and gender in the denominator model. No trimming was applied; weights ranged from 0.932 to 1.541, the effective sample size was 389.4 of 391, and no included case had a predicted inclusion probability below .10.

Table 5. Missing-data sensitivity for the secondary mediation model

Quantity	**MI**	**IPW**
Package → IMI (*a*)	0.93 (0.10) [0.72, 1.13]	0.95 (0.10) [0.73, 1.16]
IMI → delayed posttest (*b*)	0.97 (0.48) [−0.03, 1.98]	1.23 (0.45) [0.30, 2.17]
Adjusted package coefficient (*c′*)	1.93 (0.79) [0.28, 3.57]	1.54 (0.64) [0.22, 2.86]
Indirect effect (*a* × *b*)	0.90 (0.45) [−0.03, 1.84]	1.17 (0.44) [0.27, 2.07]

Note. Entries are unstandardised coefficients, with standard errors in parentheses and 95% CIs in brackets. MI = multiple imputation; IPW = inverse-probability weighting; CI = confidence interval. MI equation-path estimates were pooled after CR2 classroom-clustered estimation; IPW equation-path confidence intervals also use CR2 inference. For the MI indirect effect, the variance incorporates an approximate classroom-clustered cross-equation covariance and is treated as a sensitivity analysis.

Other checks were consistent with the focal results. Adding the 0–5 device-ownership index in the common $N = 383$ sample yielded $a = 0.93$, $b = 1.17$, an indirect estimate of 1.09, $c' = 1.63$, and a total coefficient of 2.71. Classroom fixed effects yielded a primary package estimate of $B = 2.63$ (95% CI [1.67, 3.59]) and mediation estimates of $a = 0.95$, $b = 1.31$, and $c' = 1.48$. Across 30 leave-one-classroom-out refits, no focal estimate changed sign. Design-consistent randomisation tests swapped condition labels only between the two fixed groups within each classroom and yielded $p < .001$ for H1, the mediation-sample total coefficient, and the *a* path. Because six of the 24 schools contributed two classrooms, school-clustered analyses were also conducted; estimates closely matched the classroom-clustered results: $B = 2.62$ (95% CI [1.73, 3.52]) for H1; $a = 0.95$, $b = 1.27$, and $c' = 1.51$ for mediation; and 1.21 (95% CI [0.43, 1.99]) for the school-cluster-bootstrap indirect estimate.

AUTHORS

Vasileios Daloukas is a doctoral candidate at the Department of Educational Sciences and Social Work, University of Patras Greece, and an Informatics teacher in Greek secondary education. His research focuses on technology-enhanced and game-based learning, multiplayer educational environments, educational assessment, and learning analytics. He has also developed open-source educational game applications for Moodle and WordPress.

Anthi Karatrantou is an Assistant Professor of Educational Technology at the Department of Educational Sciences and Social Work, University of Patras, Greece. Her research interests include emerging technologies in education, e-learning, STEM education, educational robotics, virtual and augmented reality, and teacher professional development.

George Koutromanos is an Associate Professor of Information and Communication Technologies in Education at the Department of Pedagogy and Primary Education, National and Kapodistrian University of Athens, Greece. His research focuses on digital game-based learning, extended reality, mobile learning, educational robotics, and technology-enhanced learning environments.

Christos Panagiotakopoulos is Professor Emeritus of Educational Technology at the Department of Educational Sciences and Social Work, University of Patras, Greece. His research focuses on the design and educational use of software, digital games, educational robotics, web technologies, and open and distance learning.