



Intraindividual Variability in Self-reported Energy as a Contextual Factor for Learning Outcomes in a K–12 Math Digital Learning Game

Marshall An¹(✉) , Mahboobeh Mehrvarz¹ , Hayden Stec¹ ,
Esma Nur Kahveci² , John Stamper¹ , and Bruce M. McLaren¹

¹ Carnegie Mellon University, Pittsburgh, PA 15213, USA

{haokanga,hstec,jstamper,bmclaren}@andrew.cmu.edu

² University of Massachusetts Dartmouth, Dartmouth, MA, USA
ekahveci@umassd.edu

Abstract. Digital learning games and intelligent tutoring systems (ITSs) have demonstrated educational benefits, yet substantial variability in learning outcomes persists. While log-based behavioral signals are often used to identify students at risk, they offer limited visibility into the fluctuating internal states that condition learning. This study examines how a log-based signal and learners' self-reported energy jointly predict immediate and delayed posttest performance (retention) in a K–12 mathematics digital learning game with an underlying ITS. The signal is operationalized as the total incorrect step-level attempts during gameplay, including repeated attempts, and self-reported energy is measured repeatedly across sessions. Data were collected from 319 students across 25 classrooms in five schools. Learning outcomes were assessed using a pretest, an immediate posttest, and a one-week delayed posttest. Despite overall significant and sustained performance improvements, incorrect attempts emerged as a consistent negative predictor of both immediate and delayed posttest performance, even after controlling for prior knowledge. Notably, intraindividual variability in self-reported energy moderated the relationship between incorrect attempts and retention: higher numbers of incorrect attempts were more strongly associated with poorer retention for learners whose self-reported energy fluctuated more across sessions. In plain terms, frequent mistakes were a stronger warning sign of weaker one-week retention for students whose self-reported energy fluctuated more across days. Robustness checks indicated that these patterns were consistent across experimental contexts. These findings suggest that intraindividual variability in self-reported energy, captured via lightweight measures, can provide contextual value for interpreting log-based signals associated with poorer retention in authentic classroom settings.

Keywords: Intelligent tutoring systems · Learning games · Self-reported energy · Intraindividual variability · Technology-enhanced learning

1 Introduction

Meta-analyses support the overall promise of digital learning games [12, 23], but outcomes are not uniform: effect sizes vary across studies and game designs [12], and even successful deployments can benefit some learners much less than others [16, 40]. One line of work in technology-enhanced learning (TEL) research addresses this variability by accommodating heterogeneous learner needs through intelligent tutoring systems (ITSs), which use techniques such as learner modeling, adaptive problem selection, and targeted feedback [3, 33, 40]. Despite ITSs’ strong record of improving average learning outcomes [30, 46], outcome variability remains pronounced. Meta-analyses report substantial heterogeneity across ITS deployments [34], and K–12 syntheses show that benefits may differ for higher- versus lower-achieving students [10, 43], motivating research that moves beyond average effectiveness to examine within-system divergence.

One established approach to understanding such variability is to analyze interaction logs for behavioral signals associated with weaker outcomes [6]. However, logs provide limited visibility into learners’ internal states that may condition how observable behaviors relate to outcomes. Attentional control and cognitive resource frameworks emphasize that learning depends on cognitive and metacognitive resources that can fluctuate over time [25, 28]. These theoretical accounts are supported by large-scale field-based evidence from K–12 educational contexts showing systematic within-day variation in students’ cognitive performance, including performance declines over the school day and partial restoration after breaks [41]. Beyond momentary performance, such fluctuations may also affect how learners monitor their understanding and regulate their responses to difficulty across sessions [5, 11, 15], suggesting a potential link between internal resource variability and retention. However, repeatedly measuring internal states in authentic classroom settings poses practical challenges, particularly when measures must be scalable and minimally disruptive [8]. To address this constraint, we introduce a lightweight, classroom-deployable self-report of learner energy administered in situ across multiple days. We use repeated energy self-reports as a pragmatic indicator of learners’ subjective experience of fluctuation across learning sessions, and examine whether this signal provides contextual value when modeled jointly with behavioral logs for predicting immediate learning outcomes and retention. We conducted the investigation using data from the Fall 2025 deployment of *Decimal Point*, a K–12 mathematics digital learning game with an underlying ITS.

This work makes two contributions to research on TEL in authentic classroom settings. First, we introduce a lightweight daily self-report energy probe suitable for in-vivo K–12 classroom use and provide initial evidence of its internal consistency and low time cost. Second, we show that intraindividual variability in self-reported energy provides contextual value for interpreting a common log-based behavioral risk signal: frequent incorrect attempts were more strongly associated with poorer retention when learners’ self-reported energy fluctuated more across days. Together, these findings contribute to understanding when

common behavioral indicators are more strongly linked to outcomes in TEL environments.

2 Related Work

2.1 Digital Learning Games with Intelligent Tutoring Systems

Prior work shows that integrating ITS components into digital learning games can support learning [3, 40]. ITSs are a well-established application of artificial intelligence in education, designed to provide adaptive, one-on-one instruction that approximates the support of a human tutor [2, 10]. Meta-analytic evidence indicates that these environments improve learning beyond typical classroom instruction, with effect sizes comparable to those of human tutors [30, 46]. However, even effective ITSs do not benefit all learners equally: learning outcomes depend both on instructional design choices and on how individual learners interact with those designs, which can vary substantially across learners [9, 32, 45]. In particular, meta-analytic work on ITSs in K–12 mathematics suggests that learning benefits are not uniform across achievement levels [43], underscoring the need to examine who benefits and under what conditions. In this context, our work moves beyond overall ITS efficacy to examine outcome variability by combining behavioral logs with self-reported learner energy.

2.2 Log-Based Signals in ITSs Associated with Outcome Variability

Prior ITS research uses interaction logs to derive signals linked to learning outcome variability, including indicators based on frequent incorrect attempts [7, 47]. Such signals are also used for real-time detection and instructional tools, such as the teacher-facing Lumilo [21], which employs error rates to help prioritize students needing assistance. Informed by this line of work, we use the frequency of incorrect attempts as a commonly available log-based indicator. Our contribution is not to propose a new behavioral metric, but to examine how this indicator’s meaning changes when interpreted alongside repeated, low-burden self-reports from authentic classroom settings.

2.3 Intraindividual Variability (IIV)

Learner energy fluctuations connect to the concept of intraindividual variability (IIV), which captures within-person changes in cognitive or behavioral performance across time and contexts that are relatively fleeting and reversible over moments or days [22, 35, 38]. Empirical work shows that IIV accounts for a substantial share of observed performance differences across populations. For example, research with older adults found that IIV accounted for approximately half of the variance in everyday cognitive performance across repeated assessments [19]. Similarly, studies of children using repeated working memory assessments in school contexts report that IIV accounts for 48%–65% of total variance

across measures [18]. Although prior IIV research has primarily examined cognitive performance, the same within-person logic can also be applied to repeated self-reports collected in classroom settings. Methodologically, within-person variability can be preferable to between-person comparisons of mean self-reports because stable individual differences (e.g., response styles or scale-use tendencies) can distort between-person differences, whereas within-person deviations capture change relative to an individual’s own typical level [14].

Accordingly, we use IIV_{energy} to denote within-learner variability in daily self-reported energy, operationalized as the standard deviation of repeated probe responses across classroom sessions. We treat this measure as a pragmatic contextual signal derived from repeated classroom self-reports, rather than as a fully validated psychological construct, and examine whether it interacts with a common log-based behavioral risk signal in relation to learning outcomes.

3 Research Questions

We formulate the following research questions: **RQ1:** How effective is the K–12 mathematics game with an underlying ITS in supporting learning overall? **RQ2:** Is a log-based behavioral indicator (number of incorrect attempts) associated with immediate posttest and delayed posttest scores, after controlling for pretest? **RQ3:** Are a learner’s mean level of self-reported energy and their IIV_{energy} associated with immediate posttest and delayed posttest scores, after controlling for pretest? **RQ4:** Do mean energy and IIV_{energy} moderate the association between incorrect attempts and immediate posttest and delayed posttest scores, after controlling for pretest?

3.1 Materials

Our study used *Decimal Point*, a digital learning game with an underlying ITS developed with CTAT [1]. As shown in Fig. 1, students “travel” through a theme park and play 24 mini-games, each with two problems, to learn decimal concepts and operations. Fantasy non-player characters (NPCs) encourage students, congratulate correct solutions, and provide feedback on mistakes.



Fig. 1. *Decimal Point* and an example mini-game.

3.2 Design and Participants

In Fall 2025, *Decimal Point* served as a shared research platform supporting two concurrent randomized experiments with distinct independent variables (one with two conditions and one with three), such that each learner was assigned to one of five platform conditions. These conditions varied in surface-level design features, including storytelling, visual theme, assessment presentation, and self-explanation orchestration. Despite different surface features, all conditions used practice and assessment items that were aligned at the item level, such that corresponding problems and test items targeted the same underlying mathematical skills, modeled as knowledge components [27]. Gameplay mechanics and logging structures were likewise held constant. The daily energy probe, central to this study, was also administered identically. Because the experimental manipulations themselves are not the focus of this study, we do not detail them here. Meanwhile, these experimental contexts provide an opportunity to examine whether observed associations are consistent across conditions, motivating the robustness checks reported in each research question. Learning outcomes were assessed with three decimal tests: pretest, immediate posttest, and one-week delayed posttest. To control for practice effects and test difficulty, three test versions (A, B, C) were counterbalanced across the six possible test sequences, with each version appearing equally often at each occasion. Each test contained 43 items, some with multiple components, for 52 possible points. To support attribution of delayed posttest performance to retention rather than additional instruction, no additional whole-class decimal instruction occurred during the study period.

A total of 481 students from 25 classrooms across five schools participated. We excluded 156 students who did not complete all required materials (pretest, immediate posttest, delayed posttest, and at least one daily energy probe). Following the outlier-exclusion procedure used in prior research with *Decimal Point* [36], we excluded 6 additional students with extreme learning gains, defined as immediate posttest minus pretest or delayed posttest minus pretest scores exceeding ± 2.5 standard deviations from the corresponding sample mean. The final analytic sample included 319 students.

Demographic data were collected via an embedded survey, with 315 participants providing demographic information. Participants were primarily in grades 5–7 (5th: $n = 101$; 6th: $n = 168$; 7th: $n = 44$), with very few outside this range (4th: $n = 1$; 8th: $n = 1$). Ages were predominantly 10–12 years (10: $n = 62$; 11: $n = 140$; 12: $n = 98$), with fewer older learners (13: $n = 13$; 14: $n = 1$); one learner reported being younger than 10. Gender was self-reported as open-ended responses; after consolidating equivalent labels, 158 participants identified as female and 151 as male. 5 responses were off-topic or invalid, and 1 participant entered “prefer not to say.” Race and ethnicity were self-reported with multiple selections allowed. 68 learners selected “I’m not sure / Prefer not to say,” and 35 selected multiple categories. Among single-category responses, the largest group identified as White or Caucasian ($n = 113$), followed by Black or African American ($n = 49$), a race/ethnicity not listed by the survey ($n = 14$),

Native American or Alaskan Native ($n = 8$), Hispanic or Latino ($n = 7$), and Asian or Pacific Islander ($n = 3$).

3.3 A Lightweight Classroom Self-report of Perceived Energy

Learners provided brief, in-situ self-reports of their perceived energy using a 6-item probe (Fig. 2) at the start of each daily session. The probe was intended as a low-burden measure for capturing day-to-day variation in self-reported energy in K–12 classrooms, rather than as a validated measure of a psychological construct. Item wording was informed by the exhaustion subscale of the Maslach Burnout Inventory for Students (MBI-SS) [39] and reworded to support brief repeated self-report in K–12 classrooms. For example, the original item “I feel tired when I get up in the morning and I have to face another day at the university” was reworded as “When I wake up for school, my brain still feels sleepy,” alongside a parallel reverse-coded item (“feels awake and ready”).

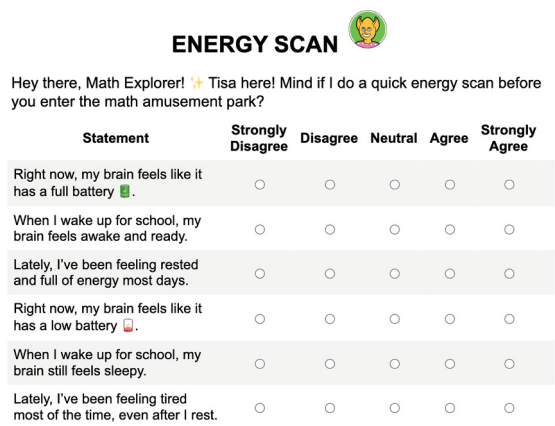


Fig. 2. The six-item daily energy probe.

This probe was designed to be lightweight and classroom-compatible, reflecting the practical constraints of an in-vivo, multi-school study conducted during regular K–12 math classes, where more intrusive or time-intensive measures (e.g., lengthy surveys or physiological sensors) were not feasible at scale. It was embedded directly into *Decimal Point* and presented automatically at the start of each session, requiring no separate materials or workflow transitions.

Consequently, the number of repeated measurements per learner was determined by the duration and instructional structure of the classroom deployment, which unfolded across the scheduled decimal class sessions over approximately one week. As a result, the present measure captures within-learner fluctuation across the available learning sessions in this instructional context, rather than a more extended, trait-like estimate of variability. Across all administrations,

participants ($N = 319$, $n = 1,810$ responses) spent an average of 30.1 s per daily session completing the probe ($SD = 30.8$ seconds). Given typical class periods ranging from approximately 45 to 85 min across participating schools, this represents well under 2% of instructional time per session.

4 Results

4.1 RQ1: Significant Overall Learning Gains from *Decimal Point*

Mean scores increased from pretest ($M = 23.04$, $SD = 10.18$) to immediate posttest ($M = 26.24$, $SD = 9.82$) and delayed posttest ($M = 27.23$, $SD = 10.15$). A one-way repeated-measures ANOVA showed a significant effect of test phase. Because Mauchly's test indicated a violation of sphericity ($W = 0.965$, $p = .003$), Greenhouse-Geisser-corrected results are reported ($\epsilon = .966$), $F(1.93, 612.94) = 105.22$, $p < .001$. Post-hoc paired t -tests with Bonferroni correction indicated significant differences between all test phases: scores improved from pretest to immediate posttest, $t(318) = 10.30$, $p_{\text{adj}} < .001$, $d_z = 0.58$, and from pretest to delayed posttest, $t(318) = 13.09$, $p_{\text{adj}} < .001$, $d_z = 0.73$, both reflecting substantial improvements. Delayed posttest scores were modestly higher than immediate posttest scores, $t(318) = 3.62$, $p_{\text{adj}} = .001$, $d_z = 0.20$. Because the Fall 2025 deployment spanned multiple experimental contexts, we examined whether overall learning outcomes differed by experiment or platform condition. Mixed ANOVA revealed no Test Phase $\times$ Experiment interaction ($F(2, 634) = 0.11$, $p = .896$) and no Test Phase $\times$ Condition interaction ($F(8, 628) = 0.40$, $p = .923$), indicating that performance improvements were consistent across contexts. Despite overall effectiveness, learner outcomes showed substantial variability. Mean learning gains were $M = 3.20$ ($SD = 5.55$) on the immediate posttest and $M = 4.19$ ($SD = 5.72$) on the delayed posttest, with standard deviations exceeding the corresponding means. This variability motivated analysis of factors underlying divergent outcomes.

4.2 RQ2: Incorrect Attempts as a Behavioral Risk Signal

RQ2 tests whether incorrect attempts, treated as a log-based behavioral risk signal, are negatively associated with immediate and delayed posttest scores. We operationalized incorrect attempts as the total number of logged step-level attempts marked incorrect during gameplay across all mini-games, using interaction logs collected through DataShop [26]. This count includes repeated attempts on the same step. Incorrect attempts were strongly right-skewed (mean = 352, median = 273; skewness = 2.07), with substantial variability across learners (IQR = 175–466; range = 17–1987). 11 learners (3.4%) made at least 1000 incorrect attempts. To examine the association between incorrect attempts and learning outcomes, we used OLS linear regression, with all numeric predictors and outcomes standardized (z-scores) to make coefficients comparable across measures on different scales. Reported coefficients therefore represent the expected

SD change in the outcome per 1 SD increase in the predictor; substantive conclusions were unchanged with unstandardized variables.

We first assessed standard OLS diagnostic checks for both models to support the use of linear regression. Multicollinearity was low based on variance inflation factors (VIFs; all VIFs < 3 , max = 1.56), and residuals showed no major departure from normality (Jarque–Bera $p = .30$ for posttest; $p = .13$ for delayed posttest). Because Breusch–Pagan tests suggested mild heteroscedasticity (posttest $p = .048$; delayed posttest $p = .018$), we report heteroscedasticity-robust (HC3) standard errors for all inferences. Controlling for prior knowledge, incorrect attempts were negatively associated with both immediate and delayed learning outcomes. For the immediate posttest, prior knowledge was a strong positive predictor ($\beta = 0.726$, $p < .001$), and incorrect attempts showed a significant negative association ($\beta = -0.202$, $p < .001$). The model explained a substantial proportion of variance in posttest performance ($R^2 = .742$, $F(2, 316) = 535.9$, $p < .001$). A parallel pattern was observed for delayed posttest performance: prior knowledge remained a strong predictor ($\beta = 0.726$, $p < .001$), and incorrect attempts continued to show a significant negative association ($\beta = -0.193$, $p < .001$). The model explained a comparable proportion of variance ($R^2 = .733$, $F(2, 316) = 471.5$, $p < .001$). Robustness checks examined whether the association between incorrect attempts and learning outcomes differed across experimental contexts. For immediate posttest performance, the association between incorrect attempts and performance was not moderated by experiment (joint Wald test, $p = .159$). However, modest variation in the strength of this association was observed across conditions nested within experiments (joint Wald test, $p = .036$). Condition-specific regressions indicated that incorrect attempts significantly predicted worse immediate performance in four of the five conditions ($p < .001$; $p = .043$; $p < .001$; $p = .007$); in the remaining condition, the effect was also negative but did not reach conventional statistical significance ($p = .082$). For delayed posttest performance, no evidence of moderation was observed, either between experiments (joint Wald test, $p = .524$) or across conditions nested within experiments (joint Wald test, $p = .876$). Overall, these results indicate a largely consistent negative association between incorrect attempts and both immediate and delayed posttest scores across experimental contexts, after accounting for prior knowledge.

4.3 RQ3: Mean Energy, IIV_{energy} , and Learning Outcomes

Before testing energy as a contextual moderator, we examined whether self-reported energy directly predicted learning outcomes. This step distinguishes models in which self-reported states have independent associations with learning from models in which they primarily condition how behavioral patterns relate to outcomes. As this daily energy probe was newly adapted for this study, we first examined its internal consistency. Across all completed responses from the 319 participants (excluding 82 incomplete responses, resulting in $N = 1,810$ responses), the six items demonstrated good internal consistency (Cronbach's $\alpha = 0.85$). The items were aggregated into a single composite energy score per

learner per daily session. Although our primary interest lies in IIV_{energy} , mean energy is included as a complementary summary of a learner’s typical reported state, allowing us to distinguish between average levels and fluctuations in self-reported energy without assuming that mean energy is itself a direct driver of learning outcomes. Accordingly, for each learner, we computed (a) mean energy as the within-learner average of daily composite scores and (b) IIV_{energy} as the within-learner standard deviation of those daily scores. To reduce the risk of unstable estimates of variability, analyses involving IIV_{energy} were restricted to learners with at least four completed daily energy probe responses, as variability estimates based on fewer observations are highly sensitive to single extreme values and may reflect measurement noise rather than meaningful within-learner variation. This criterion reduced the analytic sample in RQ3 and RQ4 from 319 to 309. As a threshold sensitivity check for computing IIV_{energy} , we repeated the RQ3–RQ4 models using alternative minimum-probe requirements (≥ 2 , ≥ 3 , and ≥ 5 completed probes). The ≥ 2 threshold corresponds to the minimum required to compute a within-learner standard deviation, while the ≥ 5 threshold corresponds to learners who completed daily energy probes on all instructional days during the one-week intervention. Results were substantively unchanged.

To test whether mean energy and IIV_{energy} are associated with immediate and delayed learning outcomes, we employed OLS regression. Consistent with RQ2, all numeric predictor and outcome variables were standardized. We estimated two models: one predicting standardized immediate posttest performance and one predicting standardized delayed posttest performance. In both models, standardized pretest score was included as a control for prior knowledge.

We assessed standard OLS diagnostic checks for both models. Multicollinearity was low, with VIFs close to 1 (maximum VIF = 1.05). Residuals showed no major departure from normality based on Jarque–Bera tests ($p = .85$ for posttest; $p = .08$ for delayed posttest). As a conservative measure, all reported inferences are based on heteroscedasticity-robust (HC3) standard errors.

For immediate posttest performance, the model explained a substantial proportion of variance ($R^2 = .707$, $F(3, 305) = 290.2$, $p < .001$). After controlling for prior knowledge ($\beta = 0.836$, $p < .001$), neither mean energy nor IIV_{energy} was significantly associated with immediate performance (both $p > .10$). For delayed posttest performance, the model similarly accounted for a large proportion of variance ($R^2 = .715$, $F(3, 305) = 277.5$, $p < .001$). After controlling for prior knowledge ($\beta = 0.838$, $p < .001$), IIV_{energy} was not significantly associated with delayed performance ($\beta = -0.014$, $p = .709$). In contrast, mean energy showed a small but statistically significant *negative* association with delayed performance ($\beta = -0.074$, $p = .016$).

A robustness check examined whether the association between mean energy and delayed posttest performance varied across experimental contexts or conditions nested within experiments. No evidence of moderation was observed by experiment (joint Wald test, $p = .266$) or by condition nested within experiment (joint Wald test, $p = .819$).

4.4 RQ4: Interaction Between IIV_{energy} and Incorrect Attempts

To examine whether self-reported energy states condition the relationship between behavioral risk indicators and outcomes, we extended the RQ3 models by adding incorrect attempts and interaction terms between incorrect attempts and (a) mean energy and (b) IIV_{energy} . As before, all continuous predictors and outcomes were standardized, pretest score controlled for prior knowledge, and heteroscedasticity-robust (HC3) standard errors were used. Multicollinearity remained low in all models (max VIF = 1.59), indicating no evidence of problematic multicollinearity with added predictors and interaction terms.

For immediate posttest performance, the interaction model explained a substantial proportion of variance ($R^2 = .740$, $F(6, 302) = 162.5$, $p < .001$). Compared with the corresponding main-effects model, adding the two interaction terms produced a small increase in explained variance ($\Delta R^2 = .006$), and the joint Wald test of the interaction terms was not statistically significant, $\chi^2(2) = 4.83$, $p = .089$. Neither the interaction between incorrect attempts and mean energy ($\beta = -0.064$, $p = .148$) nor the interaction between incorrect attempts and IIV_{energy} ($\beta = -0.047$, $p = .094$) reached conventional statistical significance, indicating no clear evidence that self-reported energy conditioned the association between incorrect attempts and immediate performance.

The key interaction effect was observed for delayed posttest performance. The interaction model accounted for a large proportion of variance ($R^2 = .750$, $F(6, 302) = 162.2$, $p < .001$). Compared with the corresponding main-effects model, adding the two interaction terms increased explained variance by $\Delta R^2 = .010$, and the joint Wald test of the interaction terms was statistically significant, $\chi^2(2) = 15.08$, $p < .001$. The interaction between incorrect attempts and IIV_{energy} was statistically significant ($\beta = -0.093$, $p < .001$), indicating that the negative association between incorrect attempts and retention was stronger for learners whose self-reported energy fluctuated more across sessions. In contrast, the interaction between incorrect attempts and mean energy was not significant ($\beta = -0.036$, $p = .345$).

A robustness check detected no differences across experiments (joint Wald test, $p = .965$) or across conditions nested within experiments (joint Wald test, $p = .778$). To evaluate whether this focal finding was robust to the nested structure of the data (students within classrooms within schools), we also re-estimated the delayed posttest interaction model using a linear mixed-effects model. The model included a random intercept for school and an additional classroom-level variance component nested within school. Because all continuous predictors and outcomes were standardized prior to analysis, the resulting coefficients are directly comparable to the OLS estimates. The mixed-effects model was fitted using restricted maximum likelihood (REML). The substantive pattern of results remained unchanged. In particular, the interaction between incorrect attempts and IIV_{energy} remained negative and statistically significant ($\beta = -0.090$, $p = .002$), closely matching the OLS estimate ($\beta = -0.093$, $p < .001$). Other coefficients were also similar in magnitude and significance.

These results suggest that the key finding was not materially altered after accounting for classroom- and school-level clustering.

Together, these findings suggest that while frequent incorrect attempts are consistently associated with poorer outcomes, IIV_{energy} amplifies this association for retention, serving as a contextual factor that helps characterize when behavioral risk indicators are more strongly linked to poorer retention.

5 Discussion

This study examined variability in learning outcomes within an overall effective K–12 mathematics digital learning game with an underlying ITS. Incorrect attempts were consistently associated with poorer immediate and retained outcomes, and IIV_{energy} amplified this association for retention. We interpret this moderation effect through a cognitive resource and attentional control perspective, in which attention and cognitive control can fluctuate with mental effort and fatigue across learning sessions [25, 28]. Field-based evidence in K–12 contexts suggests meaningful within-learner variability in cognitive performance associated with time of day, sleep quality, and tiredness [17, 29, 41]. Within this framework, IIV_{energy} may function as a contextual indicator of instability in attentional and cognitive control resources. When perceived energy is less consistent, higher frequencies of incorrect attempts may more often co-occur with more superficial processing or less consistent persistence through impasses, patterns that have been linked to poorer learning and retention in ITS contexts [7, 47]. Self-regulated learning (SRL) theory complements this account by emphasizing that effective learning depends on maintaining monitoring and regulation processes over time [4, 37, 48]. From this perspective, instability in perceived energy may disrupt these regulatory processes across sessions, strengthening the link between error-prone behavior and weaker retention.

5.1 Implications for Classroom TEL

The findings reinforce that incorrect attempts remain an informative behavioral signal, while also suggesting that repeated, lightweight self-reports can help contextualize when this common indicator is more strongly associated with poorer retention. Rather than treating log-based risk signals as having a fixed meaning across learners and sessions, the present results point to the value of interpreting such signals in relation to low-burden measures of learner state collected under authentic classroom constraints. In this sense, the contribution of the present work is less about proposing a new adaptive model and more about clarifying when a widely available behavioral indicator may be more or less meaningful in classroom TEL environments.

These findings also have implications for real-time classroom awareness and orchestration tools. Systems such as Lumilo [21] address a central challenge in authentic classrooms—supporting teachers’ help-giving decisions under time

constraints—by augmenting instructors’ real-time perceptions of students’ learning, metacognition, and behavior. The present work points to a natural extension of such systems: incorporating self-reported learner energy as a contextual signal for interpreting log-based indicators. For example, Lumilo’s “struggling” indicator incorporates recent error rates; contextualizing this signal with IIV_{energy} may yield a more nuanced picture of student need, particularly for anticipating weaker retention. Future work could prototype dashboards that surface these integrated signals and test whether they improve teachers’ support targeting.

5.2 Limitations and Future Work

Several limitations should be considered when interpreting the findings. First, our behavioral analyses focused on a single log-derived indicator (incorrect attempts). Although this measure is informed by prior work and has the strength of being commonly available across systems, which supports replication, it can arise from various mechanisms (e.g., decimal misconceptions [24, 42], slips [13], frustration [8], wheel-spinning [47], and “gaming the system” [7]). Accordingly, we interpret incorrect attempts as a behavioral risk signal rather than a diagnostic measure. Future work could examine whether additional log-based features (e.g., response timing, help-seeking behaviors, etc.) interact with IIV_{energy} in ways that further refine risk identification.

Second, learners’ behaviors and self-reported energy unfolded within socially organized classroom contexts [20, 31], which were not systematically measured. Variation in instructor monitoring and one-on-one guidance may have influenced both learners’ reported energy states and error patterns in ways our operationalizations cannot capture. In particular, a large-scale study in K–12 classrooms ($N = 1116$) suggests that the meaning and consequences of making errors depend on instructional norms surrounding help-seeking and responses to mistakes [44]. Thus, our findings characterize relationships within the instructional ecology of the Fall 2025 *Decimal Point* deployment. Future research could examine whether these patterns generalize across classrooms with different instructional norms.

Third, the daily probe captured learners’ self-reported energy rather than objective cognitive states. Because mean self-reports are sensitive to stable between-person differences [14], our primary investigation centered on within-person variability (IIV_{energy}). The observed negative association between mean self-reported energy and retention should therefore be interpreted cautiously as a secondary finding that motivates further examination.

Fourth, IIV_{energy} was estimated over a relatively short observation window spanning the available classroom sessions within the instructional unit. This design reflects the duration of the deployment and the goal of preserving classroom-feasible, low-burden measurement under authentic conditions, rather than extending the study solely to obtain more repeated observations. Thus, the present analysis shows that short-horizon fluctuation in self-reported energy can be meaningfully examined within an authentic classroom deployment, not that the measure captures a stable longer-term trait. Future work could examine whether similar patterns hold across longer instructional windows.

Fifth, delayed posttests in classroom settings reflect performance after a naturalistic interval during which learners may encounter content through unmeasured channels (e.g., individual study or incidental practice). While we ensured that no additional whole-class instruction occurred, these results represent performance under natural learning conditions rather than pure retention attributable solely to the intervention. Future work could incorporate ecologically feasible measures of intervening exposure, such as low-burden learner self-report.

Sixth, the analytic sample was nested within 25 classrooms across five schools, so classroom- and school-level context remains relevant. Teacher practices, classroom norms, and school-level factors could influence both engagement with the game and test performance. We addressed the nested structure of the data by re-estimating the focal delayed-posttest interaction model using a multilevel specification with a random intercept for school and a classroom-level variance component, which yielded the same substantive conclusion. However, the limited number of schools constrains the precision of school-level variance estimates. Future work with more schools and richer classroom-level measures could better examine contextual heterogeneity and potential cross-level moderators.

Finally, this study identifies conditions under which incorrect attempts are more strongly associated with poorer retention, but does not test whether adaptive support integrating IIV_{energy} improves outcomes. Future work should examine whether such context-rich adaptations enhance retention.

5.3 Conclusion

This study examined why learning outcomes vary across individuals even within an overall effective K–12 mathematics game with an underlying ITS. A key finding is that IIV_{energy} amplified the negative association between a behavioral risk signal (incorrect attempts) and retention: frequent errors were more strongly associated with poorer retention for learners whose self-reported energy fluctuated more across sessions. This result suggests that repeated, lightweight self-report measures can provide contextual value for interpreting log-based behavioral indicators in authentic classroom settings. More broadly, this work helps characterize when a common behavioral signal is more strongly associated with poorer retention, highlighting how low-burden classroom measures can support more context-aware interpretation of learner behavior in TEL environments.

References

1. Aleven, V., McLaren, B., Sewall, J., Koedinger, K.R.: Example-tracing tutors: a new paradigm for intelligent tutoring systems (2009)
2. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: lessons learned. *J. Learn. Sci.* 4(2), 167–207 (1995)
3. Arroyo, I., Beal, C., Murray, T., Walles, R., Woolf, B.P.: Web-based intelligent multimedia tutoring for high stakes achievement tests. In: International Conference on Intelligent Tutoring Systems, pp. 468–477. Springer (2004)

4. Azevedo, R., et al.: Lessons learned and future directions of metatutor: leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system. *Front. Psychol.* **13**, 813632 (2022)
5. Baars, M., Wijnia, L., de Bruin, A., Paas, F.: The relation between students' effort and monitoring judgments during learning: A meta-analysis. *Educ. Psychol. Rev.* **32**(4), 979–1002 (2020)
6. Baker, R.S., Corbett, A.T., Koedinger, K.R.: Detecting student misuse of intelligent tutoring systems. In: *International Conference on Intelligent Tutoring Systems*, pp. 531–540. Springer (2004)
7. Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z.: Off-task behavior in the cognitive tutor classroom: when students “game the system”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 383–390. CHI '04, Association for Computing Machinery, New York, NY, USA (2004)
8. Baker, R.S., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C.: Better to be frustrated than bored: the incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *Int. J. Hum. Comput. Stud.* **68**(4), 223–241 (2010)
9. Bouchet, F., Harley, J.M., Trevors, G.J., Azevedo, R.: Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *J. Educat. Data Min.* **5**(1), 104–146 (2013)
10. du Boulay, B.: Recent meta-reviews and meta-analyses of AIED systems. *Int. J. Artif. Intell. Educ.* **26**(1), 536–537 (2016)
11. de Bruin, A.B., Janssen, E.M., Waldeyer, J., Stebner, F.: Cognitive load and challenges in self-regulation: an introduction and reflection on the topical collection. *Educ. Psychol. Rev.* **37**(3), 1–12 (2025)
12. Clark, D.B., Tanner-Smith, E.E., Killingsworth, S.S.: Digital games, design, and learning: a systematic review and meta-analysis. *Rev. Educ. Res.* **86**(1), 79–122 (2016)
13. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User-Adap. Inter.* **4**(4), 253–278 (1994)
14. Curran, P.J., Bauer, D.J.: The disaggregation of within-person and between-person effects in longitudinal models of change. *Annu. Rev. Psychol.* **62**(1), 583–619 (2011)
15. David, L., Biwer, F., Baars, M., Wijnia, L., Paas, F., De Bruin, A.: The relation between perceived mental effort, monitoring judgments, and learning outcomes: a meta-analysis. *Educ. Psychol. Rev.* **36**(3), 66 (2024)
16. Decker-Woodrow, L.E., et al.: The impacts of three educational technologies on algebraic understanding in the context of COVID-19. *AERA Open* **9**, 23328584231165919 (2023). <https://doi.org/10.1177/23328584231165919>
17. Dewald, J.F., Meijer, A.M., Oort, F.J., Kerkhof, G.A., Bögels, S.M.: The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: a meta-analytic review. *Sleep Med. Rev.* **14**(3), 179–189 (2010)
18. Dirk, J., Schmiedek, F.: Fluctuations in elementary school children's working memory performance in the school context. *J. Educ. Psychol.* **108**(5), 722 (2016)
19. Gamaldo, A.A., Allaire, J.C.: Daily fluctuations in everyday cognition: is it meaningful? *J. Aging Health* **28**(5), 834–849 (2016)
20. Greeno, J.G., Collins, A.M., Resnick, L.B., et al.: Cognition and learning. *Handbook Educat. Psychol.* **77**, 15–46 (1996)
21. Holstein, K., Hong, G., Tegene, M., McLaren, B.M., Aleven, V.: The classroom as a dashboard: co-designing wearable cognitive augmentation for K-12 teachers. In: *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pp. 79–88 (2018)

22. Hulstsch, D.F., MacDonald, S.W., Dixon, R.A.: Variability in reaction time performance of younger and older adults. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **57**(2), P101–P115 (2002)
23. Hussein, M.H., Ow, S.H., Elaish, M.M., Jensen, E.O.: Digital game-based learning in K-12 mathematics education: a systematic literature review. *Educ. Inf. Technol.* **27**(2), 2859–2891 (2022)
24. Irwin, K.C.: Using everyday knowledge of decimals to enhance understanding. *J. Res. Math. Educ.* **32**(4), 399–420 (2001)
25. Kahneman, D.: *Attention and effort* (1973)
26. Koedinger, K.R., Baker, R.S., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the EDM community: the PSLC datashop. *Handbook Educat. Data Min.* **43**, 43–56 (2010)
27. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. *Cogn. Sci.* **36**(5), 757–798 (2012)
28. Kok, A.: Cognitive control, motivation and fatigue: a cognitive neuroscience perspective. *Brain Cogn.* **160**, 105880 (2022)
29. Könen, T., Dirk, J., Schmiedek, F.: Cognitive benefits of last night’s sleep: daily variations in children’s sleep behavior are related to working memory fluctuations. *J. Child Psychol. Psych.* **56**(2), 171–182 (2015)
30. Kulik, J.A., Fletcher, J.D.: Effectiveness of intelligent tutoring systems: a meta-analytic review. *Rev. Educ. Res.* **86**(1), 42–78 (2016)
31. Lave, J., Wenger, E.: *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press (1991)
32. Leite, W.L., et al.: Do intelligent tutoring systems benefit K-12 students? A meta-analysis and evaluation of heterogeneity of treatment effects in the US (2025). [arXiv:2511.04997](https://arxiv.org/abs/2511.04997) arXiv preprint
33. Long, Y., Aleven, V.: Gamification of joint student/system control over problem selection in a linear equation tutor. In: *International Conference on Intelligent Tutoring Systems*, pp. 378–387. Springer (2014)
34. Ma, W., Adesope, O.O., Nesbit, J.C., Liu, Q.: Intelligent tutoring systems and learning outcomes: a meta-analysis. *J. Educ. Psychol.* **106**(4), 901 (2014)
35. MacDonald, S.W., Li, S.C., Bäckman, L.: Neural underpinnings of within-person variability in cognitive functioning. *Psychol. Aging* **24**(4), 792 (2009)
36. McLaren, B.M., Adams, D.M., Mayer, R.E., Forlizzi, J.: A computer-based game that promotes mathematics learning more than a conventional approach. *Int. J. Game-Based Learn. (IJGBL)* **7**(1), 36–56 (2017)
37. Panadero, E.: A review of self-regulated learning: six models and four directions for research. *Front. Psychol.* **8**, 422 (2017)
38. Ram, N., Gerstorf, D.: Time-structured and net intraindividual variability: tools for examining the development of dynamic characteristics and processes. *Psychol. Aging* **24**(4), 778 (2009)
39. Schaufeli, W.B., Martinez, I.M., Pinto, A.M., Salanova, M., Bakker, A.B.: Burnout and engagement in university students: a cross-national study. *J. Cross Cult. Psychol.* **33**(5), 464–481 (2002)
40. Shute, V.J., Smith, G., Kuba, R., Dai, C.P., Rahimi, S., Liu, Z., Almond, R.: The design, development, and testing of learning supports for the physics playground game. *Int. J. Artif. Intell. Educ.* **31**(3), 357–379 (2021)
41. Sievertsen, H.H., Gino, F., Piovesan, M.: Cognitive fatigue influences students’ performance on standardized tests. *Proc. Natl. Acad. Sci.* **113**(10), 2621–2624 (2016)

42. Stacey, K., Helme, S., Steinle, V.: Confusions between decimals, fractions and negative numbers: a consequence of the mirror as a conceptual metaphor in three different ways. *PME Conf.* **4**, 4–217 (2001)
43. Steenbergen-Hu, S., Cooper, H.: A meta-analysis of the effectiveness of intelligent tutoring systems on K-12 students' mathematical learning. *J. Educ. Psychol.* **105**(4), 970 (2013)
44. Steuer, G., Rosentritt-Brunn, G., Dresel, M.: Dealing with errors in mathematics classrooms: structure and relevance of perceived error climate. *Contemp. Educ. Psychol.* **38**(3), 196–210 (2013)
45. Taub, M., Azevedo, R.: How does prior knowledge influence eye fixations and sequences of cognitive and metacognitive SRL processes during learning with an intelligent tutoring system? *Int. J. Artif. Intell. Educ.* **29**(1), 1–28 (2019)
46. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **46**(4), 197–221 (2011)
47. Wang, Y., Kai, S., Baker, R.S.: Early detection of wheel-spinning in ASSISTments. In: *International Conference on Artificial Intelligence in Education*, pp. 574–585. Springer (2020)
48. Zimmerman, B.J.: Becoming a self-regulated learner: an overview. *Theory Pract.* **41**(2), 64–70 (2002)