Revisiting the Hint Button: Consistent Negative Associations Between Unproductive Hint Use and Learning Outcomes in Intelligent Tutoring Systems

Marshall An
Carnegie Mellon University
Pittsburgh, PA, USA
haokanga@andrew.cmu.edu

Mahboobeh Mehrvarz
Carnegie Mellon University
Pittsburgh, PA, USA
mmehrvar@andrew.cmu.edu

John Stamper
Carnegie Mellon University
Pittsburgh, PA, USA
jstamper@andrew.cmu.edu

Bruce M. McLaren
Carnegie Mellon University
Pittsburgh, PA, USA
bmclaren@andrew.cmu.edu

Abstract

Intelligent Tutoring Systems (ITSs) commonly provide on-demand multi-level hints designed to scaffold learning, yet their relationship with learning outcomes remains complex. While unproductive hint-use behaviors are well documented, existing detection methods often rely on sophisticated models or tutor-specific features that hinder broader adoption. Through a multi-semester analysis of 999 K–12 mathematics students, this study demonstrates that simple, interpretable indicators of unproductive hint use—premature hint requests and superficial hint reading—are consistently associated with reduced learning gains across cohorts. These associations persist after controlling for prior knowledge and are particularly pronounced among learners with lower prior knowledge. While the behaviors themselves have been documented in prior literature, the primary contribution of this work lies in the robust, replicated validation of these indicators that are readily implementable across real-world educational settings. We then situate these results within an affordance perspective, arguing that the common "hint button" design in ITSs can inadvertently enable bypass strategies. Taken together, our work highlights the critical need to align hint functionality with its pedagogical purpose, ensuring that hints operate as intended scaffolds rather than shortcuts.

CCS Concepts

• **Applied computing** → **Interactive learning environments**; • **Human-centered computing** → *Human computer interaction (HCI)*.

Keywords

Learning Analytics, Intelligent Tutoring Systems, Hint Systems

ACM Reference Format:

Marshall An, Mahboobeh Mehrvarz, John Stamper, and Bruce M. McLaren. 2026. Revisiting the Hint Button: Consistent Negative Associations Between Unproductive Hint Use and Learning Outcomes in Intelligent Tutoring Systems. In *LAK26: 16th International Learning Analytics and Knowledge Conference (LAK 2026), April 27–May 01, 2026, Bergen, Norway*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3785022.3785040

LAK 2026, Bergen, Norway

ACM ISBN 979-8-4007-2066-6/26/04
https://doi.org/10.1145/3785022.3785040

1 Introduction

Intelligent Tutoring Systems (ITSs) are computer-based learning environments that provide real-time, individualized instruction and feedback, approximating the adaptive support of one-on-one human tutoring [4]. A systematic review by VanLehn has shown that ITSs can be nearly as effective as human tutoring while operating at scale [38].

One widely adopted form of adaptive support in ITSs is the provision of on-demand, multi-level hints. These hints are typically structured as a sequence that progresses from high-level, facilitative prompts toward increasingly specific guidance, culminating in a "bottom-out" hint that reveals the final answer or solution step [3]. The pedagogical theory behind this design is rooted in Vygotsky's concept of the Zone of Proximal Development [39], which posits that timely assistance can bridge the gap between a student's current ability and their potential learning. Consistent with this view, prior work suggests that students benefit more when they can regulate the timing of hint requests, receiving assistance when it is most needed rather than automatically [33]. To support this hint delivery mechanism in practice, ITS authoring tools such as Cognitive Tutor Authoring Tools (CTAT) [1, 2] provide built-in support for instructional designers to incorporate on-demand hints via a persistent hint button and a multi-level, navigable hint window—a common interface design readily available for implementation.

Despite their strong theoretical grounding, the presence of well-designed hints does not guarantee productive use. A substantial body of research shows that some learners interact with learning technologies in ways that bypass meaningful cognitive engagement, thereby diminishing learning outcomes. In the context of ITSs, such behaviors are often described as "gaming the system" [7, 10]. For example, empirical studies have shown that immediately requesting hints upon encountering difficulty is associated with poorer learning outcomes [36]. At the same time, a comprehensive synthesis of research on help-seeking in ITSs concluded that while on-demand hints may be less beneficial than previously assumed, they remain helpful under certain conditions and should therefore continue to be included in ITS design [3].

Informed by Aleven et al.'s recommendation that on-demand hints should continue to be included in ITSs [3], we incorporated on-demand, multi-level hints into our digital learning game with an underlying ITS for K–12 mathematics, *Decimal Point*. While *Decimal Point* produces consistent learning gains across semesters, inspection of pre–post gain distributions revealed substantial variability in individual learning outcomes, with standard deviations often exceeding the mean gains. This pattern suggests that while the system is effective overall, students benefit unevenly from the available instructional support. This variability motivated closer examination of how learners interact with the system, particularly how they make use of on-demand hints.

Prior work has documented a range of "gaming the system" behaviors in ITSs, along with methods for detecting them. However, many existing detection approaches rely on sophisticated machine-learned models or tutor-specific features, limiting their transferability and making them difficult to adapt to our game-based learning context. To identify feasible indicators for our setting, we leveraged fine-grained interaction logs collected from *Decimal Point*, which record every learner action. Our goal was to identify simple, interpretable, and readily adoptable indicators that are consistently associated with learning outcomes and usable by researchers across diverse educational settings. To this end, our work investigates the following research questions:

RQ1: To what extent is the digital learning game with an underlying ITS effective for K–12 mathematics learning, regardless of hint usage?

RQ2: Which interpretable indicators of unproductive hint use are associated with learning gains, and how consistently do these associations replicate across semesters?

This paper makes three contributions. First, using data from 999 students across three semesters, we provide replicated correlational evidence that two simple hint-use behaviors—premature hint requests and superficial hint reading—are consistently associated with poorer learning outcomes. These associations persist after controlling for prior knowledge and are particularly pronounced among learners with lower prior knowledge. Second, we operationalize these behaviors as lightweight, interpretable indicators that are readily implementable across real-world educational settings. Third, we use an affordance perspective to argue that the common hint-button mechanism can enable bypass strategies, motivating the redesign of hint delivery to better align with the pedagogical goal of productive struggle.

2 Related Work

2.1 Intelligent Tutoring Systems (ITSs)

ITSs represent a well-established application of artificial intelligence in education, designed to provide adaptive, one-on-one instruction that emulates the support of a human tutor [4]. These interactive learning environments have been shown to consistently improve learning beyond the outcomes of typical classroom instruction [24, 38], establishing them as a powerful tool in educational technology. The primary goal of an ITS is to foster robust learning—characterized by durable retention, efficient problem-solving, and the transfer of knowledge to new contexts [22]. A key measure of ITS effectiveness is learning gain, typically defined as the difference between pretest and posttest scores. Meta-analyses confirm that ITSs yield significant learning gains, with effect sizes comparable to those of human tutors [24, 38].

However, the effectiveness of an ITS is not guaranteed, as it depends on specific design features. ITSs differ widely in their support strategies, including proactive versus on-demand hints, feedback granularity, and meta-cognitive scaffolds [16, 26]. Consequently, comparing outcomes across systems requires close attention to these design choices. A case in point is the hint delivery mechanism. Research on help-seeking shows that the mere availability of hints is insufficient [3]. Accordingly, Aleven et al. argue that an ITS should also guide when and how students request help to realize the full benefits of tutoring [3].

Within this context, our paper contributes to ongoing research by examining how a common on-demand hint design can, under certain conditions, lead to counterproductive learner behaviors. This work aims to refine our understanding of how to optimize hint delivery in adaptive learning environments.

2.2 Hints in ITSs

Hints are a fundamental scaffolding mechanism in computer-based learning environments, designed to provide immediate, personalized support [3, 28]. Their design and utility are grounded in cognitive tutoring principles, which aim to reduce extraneous cognitive load and guide learners through complex problem-solving sequences [3, 38].

In practice, hints in ITSs are most often organized as multi-level scaffolds, beginning with facilitative or conceptual prompts and progressing through increasingly specific guidance, ultimately providing the direct answer as a "bottom-out" hint as a last resort [3, 20, 38]. This graduated design is intended to balance the need for support with the pedagogical goal of fostering independence by providing the minimal assistance necessary for the student to progress [6]. The theoretical underpinning for this approach is rooted in Vygotsky's concept of the Zone of Proximal Development [39] and the "assistance dilemma" discussion in learning sciences [20], which concerns optimizing the timing and amount of instructional support to maximize long-term learning.

Despite their theoretical benefits, empirical evidence regarding the effectiveness of hints is mixed. While controlled studies report benefits when hints are well-scaffolded, large-scale field deployments often show null or even negative associations between hint access and learning gains [27, 30, 34, 38]. This inconsistency highlights a critical gap: the need to move beyond studying hint availability to understanding how students interact with hints in real-world settings, and how specific patterns of use mediate learning outcomes.

2.3 "Gaming the System" Behaviors in ITSs

"Gaming the system" behaviors—where students exploit system features to progress without engaging in the intended cognitive work—are widely documented in Intelligent Tutoring Systems, with estimates suggesting that 10–40% of students engage in such behaviors at least intermittently [7, 8]. Typical manifestations include rapid guessing, repetitive submissions, and misuse of the hint system to obtain answers with minimal effort [8, 40]. Critically, these behaviors are associated with poorer learning gains [10, 11, 14, 36].

A substantial body of prior work in the learning analytics and educational data mining communities has focused on detecting gaming behaviors using technically sophisticated, machine-learned models. For example, Baker et al. [8] introduced a machine-learned Latent Response Model that inferred gaming using 24 log-derived behavioral features and latent-variable interaction structures, illustrating both the promise and complexity of modeling unobservable learner behaviors. Subsequent work expanded these approaches using large-scale Cognitive Tutor datasets, distilling 26 fine-grained features per student action and exploring a model space spanning approximately 10^{13} possible models [12]. Identifying effective detectors in this space often involved advanced feature selection and search techniques, such as Fast Correlation-Based Filtering and Forward Selection, as documented in [12]. Similarly, in SQL-Tutor, Baker et al. developed gaming detectors using 40 log-derived features per action [13], again underscoring the degree of feature engineering and modeling expertise required.

Despite their predictive performance, the practical adoption of these machine-learned detectors is hindered by two major barriers: the expertise required for their development and their limited generalizability. The substantial feature engineering and complex modeling techniques involved place them beyond the reach of many educational practitioners. Furthermore, building a generalizable detector requires successful transfer across student cohorts and tutor lessons—a task complicated by lesson-specific interfaces and log semantics that can make equivalent behaviors appear dissimilar across contexts [9]. Empirical evaluations show that transfer performance is often inconsistent [12], suggesting these detectors remain tightly coupled to the environments for which they were developed. Together, these constraints limit the portability of such machine-learned detectors to other ITSs and educational settings.

Alternatively, human-defined heuristics [17, 31] offer transparent and easily interpretable indicators of disengagement, making them valuable for educational practitioners. However, these heuristics are typically developed and validated within a limited number of classes or cohorts, with little evidence of replication across semesters or learner populations. As a result, their robustness and generalizability across educational contexts remain unclear.

This leaves a practical need for detection methods that are both generalizable across cohorts and simple enough for broad adoption. Our work bridges this gap by demonstrating that simple, interpretable heuristics—when rigorously validated across multiple cohorts—can provide both practical utility and scientific robustness. Specifically, we examine whether two straightforward hint-use behaviors can serve as robust indicators that generalize across cohorts, while retaining the transparency and implementability that support adoption across diverse educational contexts.

3 Methods

3.1 Materials

Our study was conducted using *Decimal Point*, a digital learning game with an underlying ITS developed with CTAT. Designed from the outset as a game, learning in *Decimal Point* is inseparable

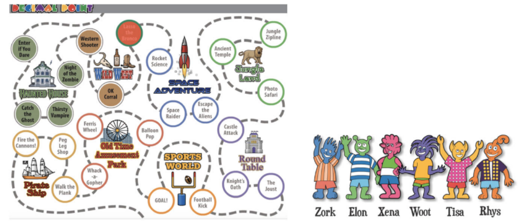
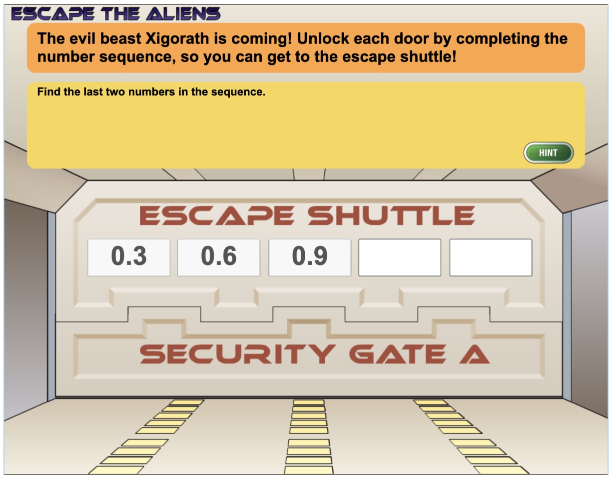


Figure 1: The Decimal Point learning game and fantasy characters that are part of the game.

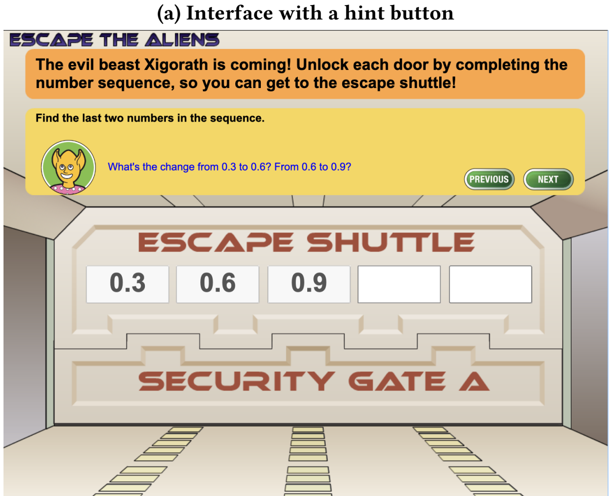
Figure 2: *Goal*, one example of a mini-game in *Decimal Point*.

from gameplay. While *Decimal Point* leverages ITS techniques to model learner behaviors and provide adaptive support, its game-first design distinguishes it from "gamified" ITSs that layer game elements such as badges, points, leaderboards, or playful agents onto otherwise conventional instructional activities while preserving tutor-like problem structures, such as MathSpring (formerly Wayang Outpost) [5], gamified SQL-Tutor [37], and gamified Lynnette [25]. In *Decimal Point*, students "travel" through a theme park playing a variety of mini-games that help them learn decimal concepts and operations, such as place value, comparing decimal magnitude, placing decimals on a number line, and adding decimals. The game features 24 mini-games and a total of 48 problems, with two problems per mini-game. Learners follow the dashed line of the amusement park map, playing mini-games in sequence, as shown in Fig. 1. A group of fantasy, non-player characters (NPCs) encourage students to play (as shown in Fig. 2), congratulate them when they correctly solve problems, and provide feedback when they make mistakes.

Fig. 3 illustrates how *Decimal Point* provides on-demand hints. Fig. 3a shows the "Hint" button; when a student clicks it, they can navigate forward and backward through three levels of hints using the "Previous" and "Next" buttons (see Fig. 3b). Fig. 3b also shows an example Level 2 hint: "What's the change from 0.3 to 0.6? From 0.6 to 0.9?" Each problem in *Decimal Point* includes three levels of hints. Level 1 hints are very general, typically reminding students of basic decimal knowledge (e.g., "First, figure out how much the numbers are changing."). Level 2 hints offer more detailed guidance on how to approach the problem (e.g., "What's the change from 0.3 to 0.6? From 0.6 to 0.9?"). Level 3 hints, also called "bottom-out hints," essentially provide the answer (e.g., "The change from one number to the next is 0.3, so the next value in the sequence is 1.2 (0.3 + 0.9)."). While the visual design is customized for the game, its hint delivery mechanism—an on-demand hint button with a



(a) Interface with a hint button



(b) Interface showing a Level 2 hint

Figure 3: *Escape the Aliens*, another example mini-game in *Decimal Point*, showcasing its multi-level hint mechanism. (a) The problem interface with a hint button. (b) An example Level 2 hint, and hint levels are navigable via "Previous" and "Next" buttons.

navigable, multi-level hint window—follows a common interface design widely used across ITSs and readily supported by CTAT.

3.2 Participants and Design

To assess learning outcomes, students completed three decimal tests (pretest, immediate posttest, and one-week delayed posttest) administered during regularly scheduled class time. A key methodological feature was the use of three counterbalanced test versions (A, B, and C). Each student was randomly assigned one of the six possible sequences of the three tests (e.g., A-B-C, B-C-A, C-A-B, etc.), ensuring that every student took each version once and that each version appeared equally often at each testing occasion. This design aimed to mitigate potential biases, attributing differences in learning outcomes to the intervention rather than to variations in assessment difficulty or practice effects from the pretest. Each test consisted of 43 items, some with multiple components, for a total possible score of 52 points.

Fine-grained learning logs were collected in DataShop [21]. Consistent with ITSs developed using CTAT, the logs capture every learner interaction with the instructional materials. These interactions can be categorized into correct attempts, incorrect attempts, and hint requests, with hint usage data serving as the primary analysis focus for RQ2.

A total of 1,370 students participated in the study, of whom (N = 999) were included in the analysis. Students were excluded if they did not complete all required materials or were identified as outliers based on a ±2.5 z-score threshold within their respective semester. Table 1 shows the number of schools and participating students every semester.

4 Results

4.1 RQ1: Significant Learning Gains from Tutor Interventions

First, we examined the overall effectiveness of the ITS intervention across the three semesters with hint access (N=999). Descriptive statistics for pretest, posttest, and delayed posttest scores are shown in Table 2. We report effect sizes (η_p^2) and interpret effects as small when $\eta_p^2 < .06$, medium when $.06 < \eta_p^2 < .14$, and large when $\eta_p^2 > .14$ [15].

A repeated-measures ANOVA revealed a significant main effect of test phase. While Mauchly's test indicated a statistically significant deviation from sphericity ($W = 0.952$, $p < .001$), the associated Greenhouse–Geisser epsilon was $\epsilon = 0.955$, indicating a negligible violation. Greenhouse–Geisser corrected results are nevertheless reported as a conservative measure, $F(1.91, 1907.18) = 124.11$, $p < .001$, $\eta_p^2 = 0.111$, with substantive conclusions unchanged from the uncorrected test.

Post-hoc pairwise comparisons with Bonferroni correction revealed significant differences between all test phases. Comparing pretest to immediate posttest scores showed a significant improvement, $F(1, 998) = 179.82$, $p < .001$, $\eta_p^2 = .153$, indicating a large effect. Comparing pretest to delayed posttest scores also revealed a significant gain, $F(1, 998) = 172.93$, $p < .001$, $\eta_p^2 = .148$, also a large effect. Although the overall mean delayed posttest score (25.30) was slightly higher than the immediate posttest score (24.88), a paired comparison indicated that this small increase was only marginally significant, $F(1, 998) = 4.17$, $p = .041$, and represented a trivial effect ($\eta_p^2 = .004$).

We also conducted separate repeated-measures ANOVAs for each individual semester (see Table 3). The pre-to-post comparisons showed consistently significant improvements across all three semesters (all $p < .001$) with effect sizes ranging from small to large ($\eta_p^2 = .059$–$.259$). Pre-to-delayed gains were also significant across semesters, though Fall 2021 showed a notably smaller effect ($\eta_p^2 = .015$, $p = .025$) compared to Spring 2021 ($\eta_p^2 = .332$) and Fall 2022 ($\eta_p^2 = .247$). Post-to-delayed effects were more variable: while Spring 2021 and Fall 2022 showed small but significant increases ($\eta_p^2 = .027$ and $.035$, $p \leq .006$), Fall 2021 showed no significant change ($\eta_p^2 = .007$, $p = .120$). As shown in Table 2, the direction of

Table 1: Number of Schools, Participants, and Final Sample by Semester

Semester	# of Schools	# of Participants	# Included in Analysis
Spring 2021	4	358	277
Fall 2021	6	436	344
Fall 2022	6	576	378

Table 2: Pretest, Immediate Posttest, and Delayed Posttest Scores by Semester

Semester	N	Pretest	Posttest	Delayed	Pre-Post	Pre-Delayed
Spring 2021	277	24.67 (10.38)	28.90 (11.25)	29.98 (11.75)	+4.23 (7.16)	+5.31 (7.55)
Fall 2021	344	20.65 (10.82)	22.32 (10.94)	21.67 (11.32)	+1.68 (6.69)	+1.02 (8.39)
Fall 2022	378	21.72 (11.32)	24.26 (11.56)	25.17 (12.12)	+2.53 (5.20)	+3.45 (6.03)
Overall	999	22.17 (11.00)	24.88 (11.55)	25.30 (12.18)	+2.71 (6.39)	+3.13 (7.52)

Note: Values are reported in the format "mean (standard deviation)".

Table 3: Repeated-measures ANOVA Results by Semester

	Pre–Post			Pre–Delayed			Post–Delayed		
Semester	F	p	η_p^2	F	p	η_p^2	F	p	η_p^2
Spring 2021	96.61	< .001	0.259	137.14	< .001	0.332	7.74	0.006	0.027
Fall 2021	21.63	< .001	0.059	5.09	0.025	0.015	2.43	0.120	0.007
Fall 2022	89.95	< .001	0.193	123.73	< .001	0.247	13.51	< .001	0.035

post-to-delayed change was positive for Spring 2021 (+1.08 points) and Fall 2022 (+0.91 points), but slightly negative for Fall 2021 (-0.66 points). These findings demonstrate consistent learning gains from the tutoring system while revealing some nuanced variability in retention patterns across different student cohorts.

Despite the consistent learning gains, the standard deviations of the gain scores were substantial, often exceeding the mean gains (see Table 2). This indicates considerable variability in individual students' benefit from the intervention, with some students showing much larger gains and others showing minimal improvement or even declines. These individual differences motivated our examination of how students' interactions with the system—particularly their help-seeking behaviors—might account for the divergent outcomes. We therefore turned to the fine-grained log data to analyze patterns of hint usage and their association with learning, which we report in the next subsection.

4.2 RQ2: Consistent Negative Associations between Unproductive Hint Use and Learning Gains Repeatedly Observed in Three Semesters

We investigated the relationship between hint usage patterns and learning outcomes by defining two metrics of unproductive hint use based on prior work. Following research on productive help-seeking behaviors [35], we operationalized unproductive hint use in two ways:

Indicator #1: Requesting hints before making any attempt. While students who are unsure of what to do may reasonably request help, evidence suggests that attempting a solution before requesting hints is more beneficial for learning, even when students feel uncertain [32].

Indicator #2: Reading hints superficially. This behavior is characterized by students advancing rapidly through hints, often skipping directly to the bottom-out hint without carefully reading earlier ones. Prior work suggests that such behavior is more strongly unproductive, as it may reflect negative affective states such as confusion, frustration, or disengagement [7].

4.2.1 Requesting hints before making any attempt. To assess the first indicator, we analyzed how often students requested a hint before attempting to solve a practice problem on their own. For each participant and problem instance, excluding tutor-performed initialization actions, we checked whether the first learner-performed action was a hint request rather than an attempt. Figure 4 shows the frequency of "hint before first attempt" in relation to pretest scores for all students.

As the visualization in Figure 4 suggests a clear trend, we quantified the relationship between hint-before-attempt actions and prior knowledge. Since both the frequency of premature hint requests and pretest scores violated normality assumptions (Shapiro-Wilk $p < 0.001$), we employed non-parametric tests. Spearman's correlation revealed a strong negative relationship between pretest scores and unproductive hint usage ($\rho = -0.567$, $p < 0.001$), indicating that students with lower prior knowledge were more likely to use hints unproductively. In addition, we divided students into pretest quartiles and found significant differences in unproductive hint usage across these groups (Kruskal-Wallis $H = 90.368$,

Table 4: Relationships Between Unproductive Hint-Use Behaviors, Prior Knowledge, and Learning Outcomes Across Three Semesters

Behavior	Semester (N)	ρ (Pre)	ρ (Post)	ρ (Delayed)	β (Post)	β (Delayed)
Premature hint requests	S21 (277)	-0.567^{***}	-0.617^{***}	-0.634^{***}	-0.28^{***}	-0.37^{***}
	F21 (344)	-0.697^{***}	-0.704^{***}	-0.694^{***}	-0.18^{***}	-0.26^{***}
	F22 (378)	-0.619^{***}	-0.616^{***}	-0.606^{***}	-0.14^{***}	-0.17^{***}
Superficial hint reading	S21 (277)	-0.710^{***}	-0.677^{***}	-0.644^{***}	-0.06^{***}	-0.08^{***}
	F21 (344)	-0.744^{***}	-0.766^{***}	-0.737^{***}	-0.11^{***}	-0.14^{***}
	F22 (378)	-0.656^{***}	-0.640^{***}	-0.636^{***}	-0.06^{***}	-0.08^{***}

Note: All Spearman correlations and OLS regression coefficients are statistically significant at $^{***}p < 0.001$.
OLS coefficients represent the relationship between the behavior and assessment scores after controlling for pretest performance.

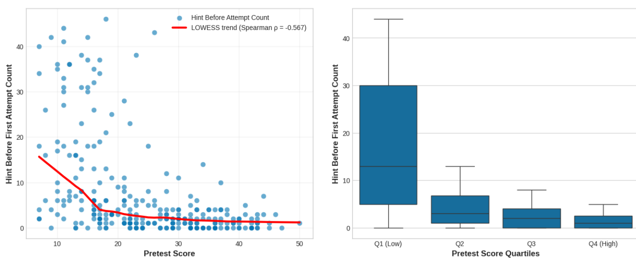


Figure 4: Relationship between prior knowledge and hint-before-attempt frequency in Spring 2021. Left: Scatter plot with a LOWESS trend line showing that hint-before-attempt frequency decreases as pretest score increases. Right: Distribution of hint-before-attempt frequency across pretest score quartiles, with higher frequencies in lower quartiles.

$p < 0.001$). As shown in Fig. 4, the scatter plot with LOWESS smoothing and the quartile-based box plots strongly demonstrate that lower-performing students tended to engage in unproductive hint use more frequently. This pattern suggests that the availability of hints may inadvertently encourage counterproductive behaviors among struggling students.

Having established the association between lower prior knowledge and more frequent hint requests before any attempts, we next examined whether this behavior is also associated with learning outcomes as measured in immediate and delayed posttests. Spearman correlations revealed that unproductive hint usage was negatively correlated with posttest scores ($\rho = -0.617$, $p < 0.001$) and delayed posttest scores ($\rho = -0.634$, $p < 0.001$). The strength of these correlations was even greater than that observed with the pretest, indicating that unproductive hint use is not only a marker of low prior knowledge students but also a meaningful behavioral indicator linked to poorer subsequent performance and reduced knowledge retention.

To complement our non-parametric correlational analysis and to estimate the relationship between unproductive hint usage and subsequent performance while controlling for prior knowledge, we conducted two ordinary least squares (OLS) regression analyses. Despite violations of univariate normality in the key variables, OLS regression is generally robust in large samples ($N = 277$), allowing estimation of associations while controlling for covariates. The first model predicted posttest scores from unproductive hint count, controlling for pretest scores. The model was significant and explained a substantial portion of the variance ($R^2 = 0.667$, $F(2, 274) = 274.9$, $p < 0.001$). First, as expected, pretest scores were a strong positive predictor of posttest scores ($\beta = 0.70$, $p < 0.001$). Second, even after controlling for prior knowledge, unproductive hint usage remained a significant negative predictor of posttest scores ($\beta = -0.28$, $p < 0.001$). A second model predicting delayed posttest scores yielded a similar pattern ($R^2 = 0.685$), with pretest scores again showing a strong positive relationship ($\beta = 0.69$, $p < 0.001$) and unproductive hint usage showing an even stronger negative coefficient ($\beta = -0.37$, $p < 0.001$). These results reinforce the findings from our correlational analysis, suggesting that unproductive hint-seeking is not merely a correlate of low prior knowledge, but shows independent associations with poorer learning outcomes and retention.

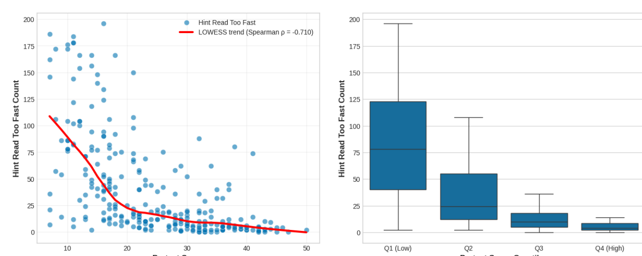


Figure 5: Relationship between prior knowledge and rapid hint reading frequency in Spring 2021. Left: Scatter plot with a LOWESS trend line showing that rapid hint reading decreases as pretest score increases. Right: Distribution of rapid hint reading frequency across pretest score quartiles, with higher frequencies in lower quartiles.

4.2.2 *Reading hints superficially.*

We similarly analyzed a second unproductive behavior, "reading hints superficially," which we identified by measuring when students moved between hints too rapidly to reasonably read them. Using an average reading speed of four words per second [19], we flagged hint sequences where the time to the next request was less than the estimated reading time for the current hint length. Both the frequency of this behavior and pretest scores violated normality assumptions, leading us to employ non-parametric tests. Spearman's correlation revealed a very strong negative relationship between pretest scores and quick hint reading ($\rho = -0.710$, $p < 0.001$), indicating that students with lower prior knowledge were significantly more likely to skim or skip hints. We then divided students into pretest quartiles and found significant differences in this behavior across the groups (Kruskal-Wallis $H = 130.144$, $p < 0.001$).

Next, we examined the relationship between quick hint reading and subsequent performance. Spearman correlations revealed that this behavior was negatively correlated with posttest scores ($\rho = -0.677$, $p < 0.001$) and delayed posttest scores ($\rho = -0.644$, $p < 0.001$), indicating it is strongly associated with poorer learning outcomes and reduced knowledge retention.

To estimate the relationship while controlling for prior knowledge, we conducted two OLS regression analyses. The first model predicted posttest scores from quick hint reading, controlling for pretest scores. The model was significant and explained a substantial portion of the variance ($R^2 = 0.659$, $F(2, 274) = 264.5$, $p < 0.001$). As expected, pretest scores were a strong positive predictor of posttest scores ($\beta = 0.67$, $p < 0.001$). Crucially, even after controlling for prior knowledge, quick hint reading remained a significant negative predictor of posttest scores ($\beta = -0.06$, $p < 0.001$). A second model predicting delayed posttest scores yielded a similar pattern ($R^2 = 0.660$), with pretest scores again showing a strong positive relationship ($\beta = 0.66$, $p < 0.001$) and quick hint reading demonstrating a significant negative coefficient ($\beta = -0.08$, $p < 0.001$). These results confirm that quickly reading hints is not merely a correlate of low prior knowledge but shows independent associations with poorer learning gains and long-term retention.

This pattern of rapid hint reading—more frequently observed among learners with lower prior knowledge, and remaining associated with poorer learning gains even after controlling for prior knowledge—is particularly concerning. It suggests that students may be strategically bypassing the learning content in earlier hints to quickly reach the bottom-out hints, which provide the final answer or direct solution steps. Students employing this gaming strategy can then simply enter the solution without engaging in the cognitive processes necessary for learning, such as understanding the problem structure, applying concepts, or developing solution strategies. This behavior effectively degrades the hint system from a scaffold for learning into a shortcut to answers, undermining its intended pedagogical benefits and potentially explaining the strong negative correlation with learning outcomes observed in our analyses.

4.2.3 Replicated Correlational Findings Across Three Semesters. The negative association between our operationalized hint-use patterns and learning outcomes has been consistently observed across three semesters (N = 277, 344, 378 respectively). As shown in Table 4, both unproductive hint behaviors demonstrated remarkably consistent patterns across all semesters, strengthening the reliability of these findings.

For premature hint requests, Spearman correlations with pretest scores ranged from $\rho = -0.567$ to -0.697 ($p < 0.001$), indicating a consistently strong negative relationship between prior knowledge and this behavior across all three cohorts. Similarly, correlations with posttest ($\rho = -0.616$ to -0.704) and delayed posttest scores ($\rho = -0.606$ to -0.694) remained strongly negative and statistically significant. More importantly, even after controlling for prior knowledge through OLS regression, premature hint requests continued to significantly negatively predict both posttest ($\beta = -0.14$ to -0.28) and delayed posttest scores ($\beta = -0.17$ to -0.37).

The pattern for superficial hint reading was similarly consistent. Correlations with pretest scores ranged from $\rho = -0.656$ to -0.744 ($p < 0.001$). The regression coefficients ($\beta = -0.06$ to -0.11 for posttest; $\beta = -0.08$ to -0.14 for delayed posttest) remained statistically significant across all semesters after controlling for pretest performance.

This consistent replication across diverse student cohorts suggests that these behavioral patterns are robust indicators of potentially at-risk students and reliable predictors of learning outcomes. Because unproductive hint use is disproportionately concentrated among lower-prior-knowledge students and remains independently associated with poorer outcomes, freely available hints may permit behavioral patterns that widen, rather than reduce, performance gaps among learners.

5 Discussion and Conclusion

This study identified two unproductive hint-use behaviors—requesting hints prematurely and reading them superficially—that were consistently and significantly correlated with reduced learning gains across multiple semesters and student cohorts. These findings contribute robust, replicated correlational evidence that refines our understanding of hint usage in digital learning environments. While the theoretical value of hints is well-established, our results underscore that the real-world efficacy of hints is mediated by how learners interact with the provided supports. The negative associations were not merely artifacts of prior knowledge; as these unproductive behaviors remained significantly negatively associated with learning gains even after controlling for pretest scores. This pattern suggests that freely available hints, in their common implementation, were associated with undesired learning strategies that bypassed the essential cognitive effort required for robust learning. These findings also address a key gap in prior work on "gaming the system" detection, which has often relied on complex, tutor-specific machine-learning models [8, 12, 13]. Although such detectors can be effective, their technical sophistication and limited transferability make them difficult to adopt across diverse educational settings. In contrast, the two indicators identified in this study are simple, interpretable, and replicated across three semesters and 999 students, demonstrating that lightweight behavioral measures can offer practical diagnostic value without requiring complex modeling pipelines.

5.1 Theoretical Interpretation of Unproductive Hint Use through the KLI Framework

These findings may be interpreted through the theoretical lens of the Knowledge-Learning-Instruction (KLI) framework [22]. The mathematical skills in *Decimal Point* primarily engage learning processes of *induction and refinement*—processes that require active construction and tuning of schemas through practice. When students rapidly access and copy bottom-out hints without engaging in meaningful problem-solving, they may circumvent the very cognitive activities that drive the acquisition of these mathematic skills. For instance, a student could solve the problem in Fig. 3 by navigating to the bottom-out hint stating "the next value is 1.2" without performing the underlying pattern recognition and calculation.

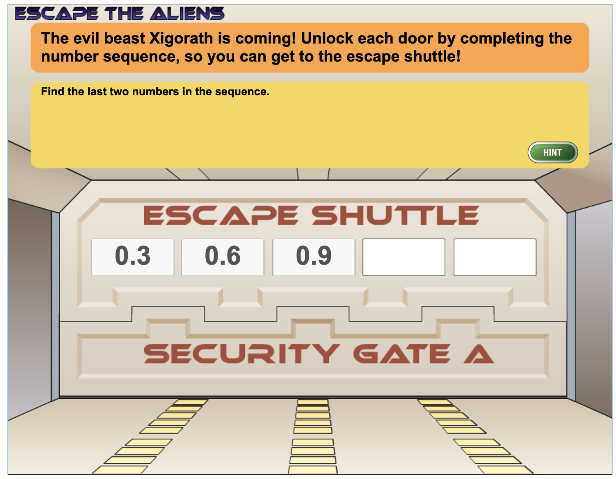
5.2 The Hint Button as an Unintended Shortcut Mechanism

The findings can be further understood by considering the hint button as a perceived affordance that permits not only intended learner interaction patterns but also undesired ones. From an affordance perspective [18, 29], the persistent visibility of the hint button may signal to certain learners that requesting help is always an available and salient action during problem solving. This affordance operates independently of—and can become misaligned with—pedagogical intent, and for low-performing students in particular, may inadvertently encourage premature reliance on external support rather than initial problem-solving effort.

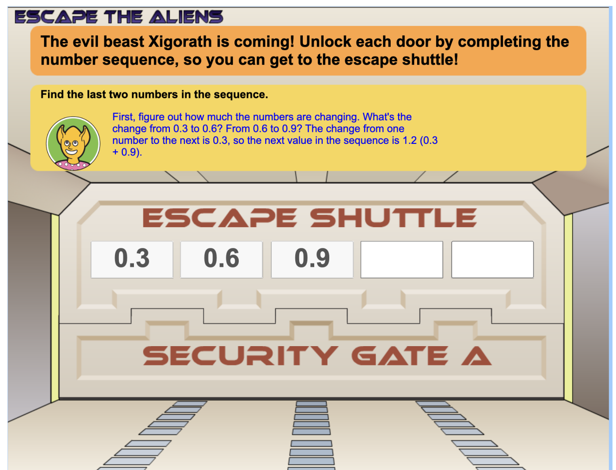
Consider a learner who, immediately upon receiving any math problem, clicks through all available hints to view the final solution. For this learner, the learning system interface with a hint button as shown in Fig. 6a effectively becomes the design shown in Fig. 6b, in which the final answer is revealed at the very beginning of every problem. This undesirably transforms the learning task into a simple copying exercise—an instructional design that is fundamentally questionable because it creates a structural misalignment with established theories of learning. The Knowledge-Learning-Instruction (KLI) framework [22] and the assistance dilemma [20] emphasize that robust learning depends on productive struggle, in which students attempt to apply knowledge before receiving help. Similarly, Vygotsky's concept of the Zone of Proximal Development [39] positions scaffolding as contingent on demonstrated effort. However, for certain learners (particularly those with low prior knowledge), the hint button offers unmediated access to final answers, thereby bypassing the very cognitive processes—schema induction and refinement—that underlie durable learning for the targeted mathematical skills.

Viewed through this lens, the two unproductive behaviors we identified are not merely instances of poor help-seeking on the learners' end. Rather, they are systematic consequences of an affordance that conflicts with pedagogical goals: premature hint requests arise because the affordance makes immediate assistance too salient, while superficial hint reading reflects the direct path the interface provides to bottom-out answers. This may explain why these behaviors replicated so consistently across cohorts and why they are disproportionately concentrated among lower-knowledge students.

Therefore, we contend that revisiting the hint delivery mechanism is not only desirable but necessary. The issue is not *whether* hints should be provided, but *how* the affordances of their delivery can be aligned with pedagogical principles so that scaffolds support productive struggle rather than enabling shortcuts to answers.



(a) The actual interface



(b) The effective interface for certain learners

Figure 6: The hint button as an affordance, exemplified using the *Escape the Aliens* mini-game from *Decimal Point.* (a) The actual interface, featuring a salient and persistently available hint button, designed with the expectation that learners will use hints appropriately and judiciously. (b) For learners who immediately seek the solution by misusing the hint button, the interface effectively collapses into this design, in which all hints—including the final answer—are revealed from the outset, transforming the task into a copying exercise.

5.3 Implications for Learning Analytics and ITS Design

A key contribution of this work is the identification and validation of simple, interpretable, and reliably detected indicators of unproductive hint usage, behaviors enabled by hint system affordances. Although demonstrated using the context of a digital learning game with an underlying ITS, these indicators are readily generalizable to a wide range of educational technologies incorporating on-demand help features. The behaviors of "premature hint requests" and "superficial hint reading" are computationally straightforward to derive from interaction logs, making them practical candidates for learning analytics dashboards or automated real-time intervention systems in diverse learning environments.

Our results necessitate a critical re-examination of a common interface design in ITSs and other adaptive learning systems: the "hint button" that, upon learner request, provides easily accessible, multi-level hints culminating in a bottom-out answer. We argue that this design can create a critical misalignment: the system's affordance of readily giving away the solution conflicts with the pedagogical goal of promoting productive struggle. We therefore contend that the primary design implication is not to remove hints, but to re-engineer hint delivery mechanisms to be more resistant to unproductive use.

Based on the findings, we propose several design directions for future research. First, it is worth investigating the efficacy of, and optimal methods for, implementing delayed hint availability—particularly for the bottom-out hint that reveals the final answer—by requiring a minimum duration of engagement or number of solution attempts before hint access is granted. This redesign directly alters the affordance that currently permits unproductive hint use, and may structurally encourage initial independent effort by removing the immediate shortcut. However, this approach primarily regulates the timing of help availability, not necessarily the quality of learners' cognitive engagement during help use. Prior work shows that interventions regulating help-seeking can improve observed help-seeking behaviors, but do not produce corresponding gains in domain-level learning, highlighting the distinction between behavioral compliance and meaningful cognitive engagement [3]. Thus, delayed access alone is unlikely to be sufficient. The core challenge extends beyond constraining when help is accessed to designing help mechanisms that elicit meaningful cognitive engagement when support is ultimately provided.

Therefore, building on Aleven et al.'s argument that hints are effective only through judicious use and cognitive engagement, and resonating with their call to explore self-explanation as a form of support [3], a second promising direction is to investigate active self-explanation as an alternative or supplement to passive hint consumption. This direction is motivated by the premise that requiring students to articulate their reasoning or reflect on observed misconceptions—instead of receiving information through passive reading, a delivery format associated with ineffective learning [23]—can better support robust learning. Historically, implementing truly interactive and dynamic tutoring support was expensive and challenging. However, in this era of large language models (LLMs), new opportunities are unlocked. We propose to explore leveraging LLMs to move beyond the static paradigm by triggering reflective, Socratic-style dialogues. These dialogues would be dynamically generated and tailored to a student's specific mistakes and inferred misconceptions, shifting the support from a static resource for passive consumption to a dynamic, conversational partner that demands active cognitive engagement. These design shifts have the potential to preserve the supportive function of hints while structurally discouraging bypass behaviors observed in our study that were consistently associated with poorer outcomes.

5.4 Limitations

Several limitations of our study should be noted. First, while our findings for RQ2 show consistent associations between unproductive hint behaviors and learning outcomes, we emphasize that these are correlational relationships from observational data and do not establish causation. The direction of these relationships and potential confounding factors should be investigated in future experimental studies. Second, we acknowledge that the behavior patterns we classified as "unproductive" may not be universally unproductive in every instance. For example, prior work has documented that some high-performing students may strategically use bottom-out hints as worked examples when encountering a novel problem type [36]. In addition, our operationalization of "superficial hint reading" relied on an estimated reading speed of four words per second [19]. While this benchmark is commonly used in psycholinguistics, actual reading rates vary across students, hint types, and contexts. Thus, some instances flagged as superficial may reflect faster-than-average but still meaningful reading. Nevertheless, the consistency of the observed correlations across cohorts suggests that the measure captures a reliable pattern of behavior that is associated with reduced learning gains at the population level. Finally, the study was conducted in the domain of K–12 mathematics. The generalizability of these specific behavioral patterns and their links to outcomes in other domains or for older learners remains an open question for future research.

5.5 Conclusion

Despite the limitations noted above, this paper provides replicated correlational evidence that unproductive hint-use behaviors—premature requests and superficial reading—are associated with worse learning outcomes in a digital learning game with an underlying ITS for K–12 mathematics. By analyzing fine-grained interaction logs across three semesters, we demonstrate that these behaviors persist as negative indicators of learning even after controlling for prior knowledge. Because these indicators are straightforward to compute from standard ITS logs and easy to interpret, they create low-overhead opportunities for replication studies and instructional interventions.

These findings challenge the community to re-examine the common design of the "hint button" in ITSs and other learning environments that provide on-demand, multi-level hints. Rather than treating hints as uniformly beneficial scaffolds, our results show that their educational value is mediated by how learners interact with them and by the affordances through which they are delivered. When hint access allows learners to bypass productive problem-solving effort, scaffolds may function counter to their pedagogical intent.

The broader implication is that research on hints should shift from *whether* to provide hints to *how* to design support mechanisms that are structurally aligned with principles of productive struggle. By grounding this argument in replicated behavioral evidence, this paper aims to inform the development of support systems in which hints operate as genuine aids to learning rather than as interactional shortcuts associated with diminished learning outcomes.

Acknowledgments

The authors used ChatGPT (OpenAI) to enhance writing clarity, then reviewed and edited the manuscript, taking full responsibility for its content.

References

[1] Vincent Aleven, Bruce M Mclaren, Jonathan Sewall, and Kenneth R Koedinger. 2009. A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education* 19, 2 (2009), 105–154.

[2] Vincent Aleven, Bruce M McLaren, Jonathan Sewall, Martin Van Velsen, Octav Popescu, Sandra Demi, Michael Ringenberg, and Kenneth R Koedinger. 2016. Example-tracing tutors: Intelligent tutor development for non-programmers. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 224–269.

[3] Vincent Aleven, Ido Roll, Bruce M McLaren, and Kenneth R Koedinger. 2016. Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 205–223.

[4] John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The journal of the learning sciences* 4, 2 (1995), 167–207.

[5] Ivon Arroyo, Carole Beal, Tom Murray, Rena Walles, and Beverly P Woolf. 2004. Web-based intelligent multimedia tutoring for high stakes achievement tests. In *International Conference on Intelligent Tutoring Systems*. Springer, 468–477.

[6] Ivon Arroyo, Kimberly Ferguson, Jeffrey Johns, Toby Dragon, Hasmik Meheranian, Don Fisher, Andrew Barto, Sridhar Mahadevan, and Beverly Park Woolf. 2007. Repairing disengagement with non-invasive interventions. In *AIED*, Vol. 2007. 195–202.

[7] Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research* 19, 2 (2008), 185–224.

[8] Ryan Shaun Baker, Albert T Corbett, and Kenneth R Koedinger. 2004. Detecting student misuse of intelligent tutoring systems. In *International conference on intelligent tutoring systems*. Springer, 531–540.

[9] Ryan Shaun Baker, Albert T Corbett, Kenneth R Koedinger, and Ido Roll. 2005. Detecting when students game the system, across tutor subjects and classroom cohorts. In *International Conference on User Modeling*. Springer, 220–224.

[10] Ryan Shaun Baker, Albert T Corbett, Kenneth R Koedinger, and Angela Z Wagner. 2004. 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*. 383–390.

[11] Ryan Shaun Baker, Ido Roll, Albert T Corbett, and Kenneth R Koedinger. 2005. Do performance goals lead students to game the system?. In *AIED*. 57–64.

[12] Ryan SJ d Baker, Albert T Corbett, Ido Roll, and Kenneth R Koedinger. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction* 18, 3 (2008), 287–314.

[13] Ryan SJ d Baker, Antonija Mitrović, and Moffat Mathews. 2010. Detecting gaming the system in constraint-based tutors. In *International conference on user modeling, adaptation, and personalization*. Springer, 267–278.

[14] Mihaela Cocea. 2009. The impact of off-task and gaming behaviors on learning: immediate or aggregate. In *Artificial Intelligence in Education*.

[15] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

[16] Benjamin Goldberg and Randall Spain. 2014. Creating the Intelligent Novice: Supporting Self-Regulated Learning and Metacognition in Educational. In *Design Recommendations for Intelligent Tutoring Systems: Volume 2-Instructional Management*. Vol. 2. US Army Research Laboratory, 109.

[17] Yue Gong, J Beck, Neil T Heffernan, and Elijah Forbes-Summers. 2010. The impact of gaming (?) on learning at the fine-grained level. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems (ITS2010) Part*, Vol. 1. 194–203.

[18] Rex Hartson. 2003. Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour & information technology* 22, 5 (2003), 315–338.

[19] Marcel Adam Just and Patricia Ann Carpenter. 1987. *The psychology of reading and language comprehension*. Allyn & Bacon.

[20] Kenneth R Koedinger and Vincent Aleven. 2007. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational psychology review* 19, 3 (2007), 239–264.

[21] Kenneth R Koedinger, Ryan SJd Baker, Kyle Cunningham, Alida Skogsholm, Brett Leber, and John Stamper. 2010. A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining* 43 (2010), 43–56.

[22] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.

[23] Kenneth R Koedinger, Jihee Kim, Julianna Zhuxin Jia, Elizabeth A McLaughlin, and Norman L Bier. 2015. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the second (2015) ACM conference on learning@ scale*. 111–120.

[24] James A Kulik and John D Fletcher. 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research* 86, 1 (2016), 42–78.

[25] Yanjin Long and Vincent Aleven. 2014. Gamification of joint student/system control over problem selection in a linear equation tutor. In *International conference on intelligent tutoring systems*. Springer, 378–387.

[26] Mehak Maniktala, Christa Cody, Tiffany Barnes, and Min Chi. 2020. Avoiding help avoidance: Using interface design changes to promote unsolicited hint usage in an intelligent tutor. *International Journal of Artificial Intelligence in Education* 30, 4 (2020), 637–667.

[27] Bruce M McLaren, J Elizabeth Richey, Huy Nguyen, and Xinying Hou. 2022. How instructional context can impact learning with educational technology: Lessons from a study with a digital learning game. *Computers & education* 178 (2022), 104366.

[28] Mahboobeh Mehrvarz, Samaneh Abdoli, Esma Nur Kahveci, Seyedahmad Rahimi, and Bruce M McLaren. 2025. The Impact of Hint Modality in Digital Game-Based Learning: A Systematic Literature Review. In *Proceedings of the 19th International Conference of the Learning Sciences-ICLS 2025, pp. 2939-2941*. International Society of the Learning Sciences.

[29] Donald A Norman. 1999. Affordance, conventions, and design. *interactions* 6, 3 (1999), 38–43.

[30] Eleanor O'Rourke, Christy Ballweber, and Zoran Popovií. 2014. Hint systems may negatively impact performance in educational games. In *Proceedings of the first ACM conference on Learning@ scale conference*. 51–60.

[31] Luc Paquette, Adriana MJB de Carvalho, and Ryan Shaun Baker. 2014. Towards Understanding Expert Coding of Student Disengagement in Online Learning. In *CogSci*.

[32] Michael Pressley, Eileen Wood, Vera E Woloshyn, Vicki Martin, Alison King, and Deborah Menke. 1992. Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational psychologist* 27, 1 (1992), 91–109.

[33] Leena Razzaq and Neil T Heffernan. 2010. Hints: is it better to give or wait to be asked?. In *International conference on intelligent tutoring systems*. Springer, 349–358.

[34] J Elizabeth Richey, Huy A Nguyen, Mahboobeh Mehrvarz, Nicole Else-Quest, Ivon Arroyo, Ryan S Baker, Hayden Stec, Jessica Hammer, and Bruce M McLaren. 2024. Understanding gender effects in game-based learning: the role of self-explanation. In *International Conference on Artificial Intelligence in Education*. Springer, 206–219.

[35] Ido Roll, Vincent Aleven, Bruce M McLaren, and Kenneth R Koedinger. 2011. Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and instruction* 21, 2 (2011), 267–280.

[36] Benjamin Shih, Kenneth R Koedinger, and Richard Scheines. 2008. A Response Time Model For Bottom-Out Hints as Worked Examples. *EDM* 2008 (2008), 117–126.

[37] Faiza Tahir, Antonija Mitrovic, and Valerie Sotardi. 2022. Investigating the causal relationships between badges and learning outcomes in SQL-Tutor. *Research and Practice in Technology Enhanced Learning* 17, 1 (2022), 7.

[38] Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist* 46, 4 (2011), 197–221.

[39] Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Vol. 86. Harvard university press.

[40] Jason A Walonoski and Neil T Heffernan. 2006. Prevention of off-task gaming behavior in intelligent tutoring systems. In *International conference on intelligent tutoring systems*. Springer, 722–724.