When the Finish Line Moves: Immediate Corrective Practice and Implications for Design in Digital Math Learning *

Wen-Chiang Lim
Worcester Polytechnic Institute
wlim@wpi.edu

Yanping Pei
Worcester Polytechnic Institute
ypei@wpi.edu

Adam Sales
Worcester Polytechnic Institute
asales@wpi.edu

Neil Heffernan
Worcester Polytechnic Institute
nth@wpi.edu

ABSTRACT

Digital learning platforms often give students a chance to solve a similar problem immediately after an incorrect response, a design we refer to here as "*immediate corrective practice*" (ICP). While prior work suggests that immediate feedback and targeted retries can support learning, less is known about how various implementations of ICP affects performance and engagement in real-world settings.

We study ICP using extant large-scale log data from a K–12 mathematics platform where teachers can choose whether to enable a feature that presents supplementary similar problems after errors on eligible items within an assignment. Focusing on students who answered the first eligible problem in an assignment incorrectly, we compare those shown ICP with those assigned the same problem sets but not shown ICP because their teachers did not use the feature. Students who encountered ICP performed better on the next problem, but were also less likely to complete the assignment.

One possible explanation is that additional practice after an error shifts students' perceived endpoint of the assignment, creating a moving-goalpost effect that discourages continued engagement. As large language models make it easier to generate corrective problems and feedback at scale, our findings offer design insights for supporting learning while minimizing disengagement.

Keywords

immediate feedback, immediate corrective practice, observational data

*(Does NOT produce the permission block, copyright information nor page numbering). For use with edm_article.cls.

Wen Chiang Lim, Yanping Pei, Adam Sales, and Neil Heffernan. When the Finish Line Moves: Immediate Corrective Practice and Implications for Design in Digital Math Learning. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 808–812. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21039687

1. INTRODUCTION

Digital learning platforms in K–12 mathematics often provide students with immediate feedback on auto-gradable, non-open-response problems. This can take the form of correctness feedback, hints, or a worked explanation. To reinforce understanding, students may also be given opportunities for "*immediate corrective practice*" (ICP), such as solving similar problems they had difficulty with, either immediately after the current problem or after completing the assignment. Examples include "Start Over" feature in Khan Academy [9], "Redo" in ASSISTments [15], and "Quick Retake" in ALEKS [1]. While prior work suggests that immediate feedback and targeted retries can support learning [14, 4], less is known about how different implementations affect student learning and engagement in real-world settings.

This paper studies one particular implementation of ICP in a real-world DLP context. Specifically, we examine the effects of providing students with a similar supplementary problem immediately after an incorrect response within an ongoing assignment. We focus on a naturally occurring setting in a K–12 mathematics platform in which teachers can enable or disable an ICP feature that presents students with a similar problem immediately after they answer a problem incorrectly on the first attempt, provided a similar problem is available. We investigate the following research questions:

- **RQ1**: Does immediate corrective practice improve student performance on the immediate next problem?
- **RQ2**: Does immediate corrective practice affect student completion of the assignment?
- **RQ3**: If the evidence suggests "Yes" to RQ2, is the decrease in persistence purely due to longer assignment length?

2. RELATED WORK

2.1 Benefits of Immediate Feedback and ICP

Immediate corrective practice is closely related to research on feedback and mastery learning. Mastery learning frameworks emphasize iterative cycles of assessment, feedback, and corrective opportunities, where students are given additional chances to practice until proficiency is reached [3, 5]. Similarly, digital learning platforms often provide immediate feedback and targeted retry opportunities after errors, and such approaches have been associated with improved learning, especially for weaker students [2, 11, 14].

The instructional logic is straightforward: when a student makes an error, the platform uses that moment to provide a closely related opportunity to apply what was learned from the immediate feedback. This is especially plausible for procedural mathematics, which benefits from repeated practice.

2.2 Potential Tradeoffs for ICP

Implementations of immediate corrective practice vary across platforms. Students may practice a fixed skill until they reach mastery, practice on similar problems after an assignment, or receive supplementary similar problems woven directly into the assignment they are working on.

Although adding similar problems after an error is consistent with mastery learning, it may not always be beneficial from the learner's perspective. Supplementary ICP within an assignment can provide a timely opportunity to revisit a skill, but it also increases the amount of work. Prior work has shown that assignment size matters for student performance, with evidence that shorter assignments can be associated with better performance than longer ones [12]. This raises the possibility of assignment fatigue, which may reduce students' willingness to complete the assignment.

In addition, research on goal-distance theory [10, 7] suggests that motivation depends on perceived distance to completion. When ICP problems are added after an error, students may perceive the assignment's endpoint as shifting after they have already started the assignment. Thus, while ICP is aligned with mastery learning, its implementation within an assignment may also lower persistence by making the finish line feel farther away than expected.

2.3 Implications for scaling of ICP

Prior to the advent of large language models (LLMs), creating similar problems often required substantial human labor, alongside work on crowdsourcing and learner-sourcing to produce high-quality educational content. Recent research suggests that LLMs can now generate and scale such problems with much less effort [6, 13]. As a result, it is increasingly important to move beyond content quality alone and examine existing implementations to identify better designs for future randomized controlled trials and field testing.

3. DATA AND SETTING

This study used extant de-identified log data collected by a K–12 mathematics digital learning platform, ASSISTments, between Jan 2023 and Dec 2025 [8], with the data collection process approved by IRB. The platform has an ICP feature that, when enabled for an assignment, provides students with a similar problem after an incorrect response on eligible items. We focus on data with these criteria:

- a student attempts the first item in an assignment for which a corresponding similar corrective practice problem is available, even though this is not necessarily the first item in the assignment,
- the student answers this item incorrectly on the first attempt, and
- there is a later non-open-response autogradable item.

These instances represent cases in which immediate corrective practice could potentially be triggered while still allowing measurable downstream outcomes without contamination from earlier ICP exposure. Our two analytic samples consist of student-problem observations meeting these criteria: a *first-instance* dataset with students' first exposure to the feature, and a *full* dataset with all such exposures. We compare outcomes between students exposed to immediate corrective practice and students assigned the same problem sets but not exposed because their teachers did not enable the feature. Assignments with no later auto-gradable item after the ICP-eligible item are excluded.

Finally, to examine whether lower assignment completion is simply due to longer assignments under ICP, we analyzed an *expanded* dataset containing all student assignment attempts between Jan 2023 and December 2025. Using this dataset, we estimated the association between teacher-enabled ICP and assignment completion while controlling for the maximum possible assignment length if ICP were triggered. We use the maximum possible length because ICP is triggered only after incorrect responses, and not all students will answer all ICP-eligible problems incorrectly.

4. MEASURES

Outcome variables. We consider the following outcomes on the immediate next autogradable problem, using the average when the problem has more than one part, as well as completion of the rest of the assignment.

Average next-problem score (RQ 1,2 only): We analyzed these scores in two ways: first, among only students who completed the next problem, and second, with students who skipped the next problem coded as 0. Scoring reflected both accuracy and support use: first-attempt correct responses earned 100%, correct responses after hints or within the second or third attempts earned partial credit (33–66%), and responses requiring more than three attempts or revealed-answer responses earned 0%.

Assignment completion (RQ 1,2,3): Whether the student goes on to complete the assignment, i.e., a binary variable.

Covariates. To control for students' general ability and current performance in the relevant skill, we include the following variables centered on the median:

Prior 5-problem average performance (RQ 1,2,3): student's average performance on the previous five problems before the assignment, used as a proxy for general ability

Initial-item score (RQ 1,2 only): score on the initial item that was answered incorrectly on first attempt, or average score if the problem had more than one part.

Assignment length (RQ3 only): Natural log of the base number of problems if ICP is not enabled, and theoretical maximum number of problems, including ICP, if it is enabled.

4.1 Analytic Approach

We estimated the association between immediate corrective practice and subsequent student outcomes using multilevel regression models fit separately in RQ1 and RQ2 on two datasets: a *first-instance* dataset and a *full* dataset. This allowed us to examine whether results were consistent when focusing on a student's first eligible exposure versus all eligible exposures. Our main predictor for RQ1 and RQ2 was `ICP_seen`, an indicator of whether a student received an ICP problem after an incorrect response.

In RQ3, the main predictor was `use_ICP`, an indicator of whether the teacher enabled the ICP feature for the assignment, and analyzed using the *expanded* dataset. The models include random intercepts for students and teachers, as well as for problems in RQ1 and RQ2 and problem sets in RQ3, to acccount for differences across these units that are not captured by the observed covariates.

4.2 Models for Continuous Outcomes

For next-problem outcomes, we fit linear mixed-effects models using `lmer` from `R`'s `lme4` package with maximum likelihood (`REML = FALSE`).

$$\begin{aligned} Y_{ijkm} = \beta_0 &+ \beta_1 \texttt{ICP\_seen}_{ijkm} \\ &+ \beta_2 \texttt{prior5avg}_{ijkm} \\ &+ \beta_3 \texttt{prior5avg}_{ijkm} * \texttt{ICP\_seen}_{ijkm} \\ &+ \beta_4 \texttt{initial\_score}_{ijkm} \\ &+ u_i + v_j + w_k + \varepsilon_{ijkm} \end{aligned}$$

where Y_{ijkm} denotes the continuous outcome for observation m. Here, u_i, v_j, and w_k denote random intercepts for student, teacher, and next problem respectively, and assumed to be normally distributed with mean 0. Residual error ε_{ijkm} is also assumed to be normally distributed with mean 0.

We applied this specification to the two continuous outcomes: next-problem score average, and its corresponding variant in which students who did not complete the next problem were scored as zero. This allowed us to test whether the observed associations were robust to a more conservative coding of non-completion.

4.3 Models for Binary Assignment Completion

For the binary outcome indicating whether the student completed the remainder of the assignment in RQ1 and RQ2, we fit a generalized linear mixed-effects model with a logit link using `glmer` from `lme4` in `R`. As in the linear models, we included random intercepts for student, teacher, and next problem, with the same distributional assumptions.

$$\begin{aligned} \text{logit}\left(Pr(Y_{ijkm} = 1)\right) = \beta_0 &+ \beta_1 \texttt{ICP\_seen}_{ijkm} \\ &+ \beta_2 \texttt{prior5avg}_{ijkm} \\ &+ \beta_3 \texttt{prior5avg}_{ijkm} * \texttt{ICP\_seen}_{ijkm} \\ &+ \beta_4 \texttt{initial\_score}_{ijkm} \\ &+ u_i + v_j + w_k \end{aligned}$$

where $Y_{ijkm} = 1$ indicates assignment completion. For interpretation, fixed effects from the logistic models were exponentiated and reported as odds ratios. In RQ3, the model is similar, except that we replace `ICP_seen` and `initial_score` with `use_ICP` and `assignment_length` respectively. Due to the size of the *expanded* dataset ($>$ 4 million records), we used `glmmTMB` in `R` which is more efficient for large datasets.

4.4 Bootstrap Inference

Given the complexity of the mixed-effects structure, we used parametric bootstrap resampling with `bootMer` to obtain empirical standard errors for the fixed effects in the models for RQ1 and RQ2. For each fitted model, we generated 500 bootstrap replications by simulating new response values from the fitted model and refitting the model to each simulated dataset, using `type = "parametric"` and `use.u = FALSE`. Bootstrap standard errors were then calculated as the standard deviation of the bootstrapped fixed-effect estimates across replications. However, we did not use bootstrapping for RQ3 because of the dataset size and computational cost of repeatedly refitting the model. Given the much larger sample size in RQ3, we do not expect the substantive conclusions to be sensitive to this choice.

5. RESULTS

Our results suggest a mixed pattern. On one hand, immediate corrective practice appears to provide short-term benefits for students who engage with it. For RQ1, students exposed to ICP show better subsequent performance, consistent with the idea that immediate targeted reinforcement can support learning after an error. This pattern holds even after accounting for students who skipped the next problem.

On the other hand, these benefits come with important engagement trade-offs. For RQ2, exposure to ICP is associated with lower assignment completion. In other words, while ICP may support learning after an error, the additional problems may also discourage students from finishing the assignment.

One way to interpret these results is to distinguish between the *instructional value* of the feature and its *psychological experience*. The same intervention that helps one student consolidate learning may be experienced by another as an unexpected increase in workload. In RQ3, we find that even after controlling for assignment length, ICP remains associated with lower odds of assignment completion. This suggests that lower completion is not due to assignment length alone, and that a perceived shift in the assignment endpoint, consistent with goal-distance theory, may also play a role. *We elaborate on the results in greater detail below.*

Across both the first-instance and full datasets, results were highly consistent. Students who saw the similar next problem (ICP) performed better on the immediate next problem, and this pattern held even when non-responses were recoded as 0 (Table 1). The ICP × Prior5avg interaction was negative across specifications, suggesting that the performance benefit of ICP was smaller for students with higher prior performance.

At the same time, students who saw ICP had lower odds of completing the assignment in both the first-instance and full

Table 1: Next Problem Score (Inclusive of non-responses coded as 0) Mixed Effects Linear Regression Results

	First-Instance	Full
(Intercept)	0.467***	0.470***
	[0.452, 0.481]	[0.458, 0.481]
ICP seen	0.036***	0.029***
	[0.023, 0.050]	[0.021, 0.036]
Prior5avg	0.248***	0.241***
	[0.237, 0.258]	[0.233, 0.248]
Initial score	0.295***	0.284***
	[0.285, 0.304]	[0.277, 0.291]
ICP seen × Prior5avg	-0.032*	-0.040***
	[-0.062, -0.001]	[-0.056, -0.023]
Total Observations	92,875	161,625
ICP observations	8,974	29,856
Unique Students	39,972	50,482
Unique Teachers	1,481	1,605
Unique Next Problems	858	1,279

Note. Entries are coefficient estimates with statistical significance indicated: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$., with 95% confidence intervals in brackets below. The interaction term shows how the association between prior 5-problem average and next-problem score differs when ICP was seen.

Table 2: Student Completed Assignment Mixed Effects Logistic Regression Results

	First-Instance	Full
(Intercept)	6.94***	6.89***
	[6.05, 7.97]	[6.11, 7.77]
ICP seen	0.82**	0.84***
	[0.72, 0.94]	[0.79, 0.90]
Prior5avg	0.93	0.89**
	[0.86, 1.01]	[0.83, 0.95]
Initial score	1.21***	1.22***
	[1.11, 1.31]	[1.15, 1.30]
ICP seen × Prior5avg	1.42**	1.16*
	[1.11, 1.81]	[1.01, 1.34]

Note. Entries are odds ratios with statistical significance indicated: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, with 95% confidence intervals in brackets below. Odds ratios above 1 indicate higher odds of assignment completion, while odds ratios below 1 indicate lower odds.

datasets (Table 2). This negative association was somewhat weaker for students with higher prior performance.

In the analysis using the expanded dataset, ICP remained associated with lower completion even after controlling for median assignment length (Table 3) and student performance. The interaction results also point in the similar directions across specifications: the negative association between ICP and assignment completion appears weaker for students with stronger prior performance and stronger for lower-performing students. However, this moderation was estimated more precisely when prior performance was modeled continuously than when it was coded as a lowest-quartile indicator. Given that lower completion is not explained by assignment length alone, it lends credibility that moving the goalpost has led to decreased persistence.

Table 3: [Expanded Dataset] Assignment Completion Mixed Effects Logistic Regression Results

	Prior5Avg	LowPrior5Avg
(Intercept)	6.49***	6.50***
	[6.13, 6.87]	[6.13, 6.88]
Use ICP	0.88***	0.88***
	[0.86, 0.90]	[0.86, 0.90]
Asg length	0.41***	0.41***
	[0.40, 0.41]	[0.40, 0.41]
Prior5avg	1.10***	
	[1.08, 1.12]	
Low_Prior5avg		0.96***
		[0.95, 0.97]
Use ICP × Asg length	1.03***	1.03***
	[1.02, 1.05]	[1.02, 1.05]
Use ICP × Prior5avg	1.04*	
	[1.01, 1.08]	
Use ICP × Low_Prior5avg		0.98
		[0.96, 1.00]
Total Observations		4,154,258
Unique Students		204,175
Unique Teachers		3,586
Unique Problems Sets		85,329

Note. Entries are odds ratios with statistical significance indicated: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, with 95% confidence intervals in brackets below. Odds ratios above 1 indicate higher odds of assignment completion, while odds ratios below 1 indicate lower odds.

6. DISCUSSION, LIMITATIONS, AND DESIGN IMPLICATIONS

This study is limited by its observational design, so the results should be interpreted as associations rather than causal effects. Because the sample was restricted to students who reached the first eligible problem, we cannot rule out selection or survival bias. Our outcomes also focus on short-term performance and assignment completion, rather than longer-term learning, and the findings are specific to the platform studied.

Within these limits, the results suggest a consistent trade-off: immediate corrective practice within an assignment is associated with slightly better next-problem performance, but also with lower assignment completion. This suggests that the value of ICP depends not only on the quality of the added problems, but also on how they are implemented. For platform developers, the key question is not simply whether to provide similar problems after errors, but how to design their use in ways that support learning without reducing persistence. This also points to promising directions for future randomized experiments, such as testing alternative implementations that vary how workload changes during an assignment, including reverse moving-goalpost designs that remove problems when students answer correctly. As large language models make it easier to generate similar problems at scale, the main challenge may shift from producing such content to designing better ways to use it.

7. REFERENCES

[1] ALEKS Corporation. Assignment suite update, 2013. https://www.aleks.com/highered/math/Assignment_

Suite_Update.pdf. Accessed: April 6, 2026.

[2] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier. Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, 4(2):167–207, Apr. 1995.

[3] B. S. Bloom. Learning for mastery. *Evaluation Comment*, 1(2):1–12, 1968.

[4] G. Campitelli and F. Gobet. Deliberate Practice: Necessary But Not Sufficient. *Current Directions in Psychological Science*, 20(5):280–285, Oct. 2011.

[5] T. R. Guskey and S. L. Gates. Synthesis of research on the effects of mastery learning in elementary and secondary classrooms. *Educational Leadership*, 43(8):73–80, 1986.

[6] J. He-Yueya, N. D. Goodman, and E. Brunskill. Evaluating and optimizing educational content with large language model judgments. In B. PaaÃŸen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 68–82, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[7] C. Heath, R. P. Larrick, and G. Wu. Goals as reference points. *Cognitive Psychology*, 38(1):79–109, 1999.

[8] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, Dec. 2014.

[9] Khan Academy. Can a student restart an exercise, quiz, or unit test? https://support.khanacademy.org/hc/en-us/articles/35941807811213-Can-a-student-restart-an-exercise-quiz-or-unit-test. Accessed: March 31, 2026.

[10] R. Kivetz, O. Urminsky, and Y. Zheng. The Goal-Gradient Hypothesis Resurrected: Purchase Acceleration, Illusionary Goal Progress, and Customer Retention. *Journal of Marketing Research*, 43(1):39–58, Feb. 2006.

[11] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark. Intelligent Tutoring Goes To School in the Big City. *International Journal of Artificial Intelligence in Education*, 8:30–43, 1997. Artwork Size: 165322 Bytes.

[12] S. Mojarad. Studying assignment size and student performance using propensity score matching. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*, pages 701–702, 2016.

[13] K. A. Norberg, H. Almoubayyed, L. De Ley, A. Murphy, K. Weldon, and S. Ritter. Rewriting Content with GPT-4 to Support Emerging Readers in Adaptive Mathematics Software. *International Journal of Artificial Intelligence in Education*, 35(2):587–626, June 2025.

[14] R. Razzaq, K. S. Ostrow, and N. T. Heffernan. Effect of Immediate Feedback on Math Achievement at the High School Level. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, editors, *Artificial Intelligence in Education*, volume 12164, pages 263–267. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science.

[15] The ASSISTments Foundation. Redo feature, 2025. https://www.assistments.org/academy-videos/redo-feature. Accessed: March 31, 2026.

8. ACKNOWLEDGMENTS

Neil and Cristina Heffernan thank their many past and current funders including NSF (2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 175 9229, 1724889, 1636782, 1535428, 2215842, 2225091, 2341948, & 2153481); IES (R305N210049, R305D210031, R305A17013 7, R305A170243, R305A180401, R305A12 0125, R305R22001 2, R305T240029, R305J250058); DOE (U411B190024, S411A 240012, S411A240012); GAANN (P200A120238, P200A1800 88), Schmidt Futures, BMGF, CZI, Arnold, Hewlett, the Jaffe Foundation, and anonymous donors. None of the opinions expressed here are those of the funders.