Evaluating the Effects of Assignment Report Usage on Student Outcomes in an Intelligent Tutoring System: A Randomized-Encouragement Design

Wen Chiang Lim
Worcester
Polytechnic Institute
Worcester, MA, USA
wlim@wpi.edu

Neil T. Heffernan
Worcester
Polytechnic Institute
Worcester, MA, USA
nth@wpi.edu

Adam Sales
Worcester
Polytechnic Institute
Worcester, MA, USA
asales@wpi.edu

As online learning platforms become more popular and deeply integrated into education, understanding their effectiveness and what drives that effectiveness becomes increasingly important. While there is extensive prior research illustrating the benefits of intelligent tutoring systems (ITS) for student learning, there is comparatively less focus on how teachers' use of ITS impacts student outcomes. Much existing research on teachers' ITS usage relies on qualitative studies, small-scale experiments, or survey data, making it difficult to identify the causal effects of their engagement with these systems.

To bridge this gap, we conducted a study using a randomized encouragement design on an online mathematics platform, where teachers were randomly assigned to one of two groups: an encouragement group or a control group. Teachers in the encouragement group received a popup prompt urging them to explore the assignment report after they created an assignment, while those in the control group did not receive any additional prompts. The study focused exclusively on teachers new to the platform, as this group was expected to be most influenced by the encouragement prompt.

The findings show that viewing the assignment report did not significantly impact the percentage of students who started the next assignment or their value-added scores. However, it did lead to a notable increase in the percentage of students completing the next assignment. This effect, confirmed using the Anderson-Rubin test (which is robust against weak instruments), demonstrates a measurable causal relationship between teachers' use of assignment reports and student outcomes. Based on data from 330 teachers, this large-scale study sheds light on the causal effects of teachers engaging with ITS data on student learning and adds to the growing evidence base for effective teaching strategies in online learning environments.

The pre-registration for the paper is available at https://osf.io/5u2n3/?view_only=39c4416ed9c04666885873b82c23f734, while data and code are available at https://osf.io/4nqxu/?view_only=1dcf5157005c4f82b815dad1fc67514a.

Keywords: randomized encouragement design, instrumental variable, intelligent tutoring systems, teaching practices

1. Introduction

Over the past decade, teachers have become more open and better prepared to use intelligent tutoring systems (ITS) in their classrooms. The enforced shift to online learning during the COVID19 pandemic has only served to further accelerate the adoption and openness towards these systems (OECD, 2023; Martin et al., 2020; Pantelimon et al., 2021). ITS have been widely recognized for their ability to enhance student learning by providing on-demand feedback, data-driven recommendations, and adaptive problem-solving support. There are also robust large-scale randomized controlled trials involving schools, teachers and students that have demonstrated significant learning gains for students using these systems (Feng et al., 2023; Pane et al., 2014; Koedinger et al., 1997)

While ITS research has largely focused on student outcomes, less attention has been given to how teachers engage with ITS data and whether this engagement influences student learning. Many ITS platforms provide teachers with detailed learning analytics reports (ASSISTments, 2024; Carnegie Learning, 2025; McGraw Hill Education, 2025) that can inform instruction, yet it remains unclear how frequently teachers use these reports or whether their use of these reports leads to measurable improvements in student outcomes.

This study seeks to fill this gap by investigating the causal impact of teachers' use of ITS assignment reports on student outcomes. Using a randomized encouragement design, we examine whether teachers can be nudged to engage with these reports and, if so, whether this engagement improves student participation and performance. Specifically, we seek to answer the following research questions:

- **Research Question 1 (RQ1):** Given the theoretical pedagogical benefits to teachers leveraging the learning analytics from the assignment report, can teachers be encouraged to view the assignment report?
- **Research Question 2 (RQ2):** Is there a statistically significant quantifiable causal impact of teachers viewing the assignment report on student outcomes? In particular, we look at the impact of the teachers looking at the first assignment report on:
  - (**RQ2a1**) the proportion of students that *started* the second assignment (teacher-level analysis), capturing the aggregated effect of a teacher's report viewing on their entire class, and (**RQ2a2**) the probability of a student *starting* the second assignment (student-level analysis), investigating whether an individual student is more likely to start the assignment based on their teacher's report viewing;
  - (**RQ2b1**) the proportion of students that *completed* the second assignment (teacher-level analysis), assessing whether a teacher's engagement with reports influences class-wide completion rates, and (**RQ2b2**) the probability of a student *completing* the second assignment (student-level analysis), investigating whether an individual student is more likely to complete the assignment based on their teacher's report viewing;
  - (**RQ2c**) value-added scores of the students in the second assignment (student-level analysis), measuring how a teacher's report viewing affects individual student performance beyond just participation rates.

Through this study, we seek to use large-scale real world data from many teachers, to investigate the impact of reviewing assignment reports, and hopefully better understand a teacher's role and contribution to how ITS has benefitted students overall as found in the large-scale randomized controlled trial studies in Feng et al. (2023) and Pane et al. (2014).

2. Background

2.1. The Role and Benefits of Intelligent Tutoring Systems

ITS are widely recognized for the capability to enhance student learning through on-demand feedback, data-driven recommendations, and adaptive problem-solving support. Research highlights the positive impact of features such as immediate feedback (Patikorn and Heffernan, 2020; Razzaq and Heffernan, 2009) and personalized support (Kochmar et al., 2020; Zhou et al., 2020; Lee et al., 2024). Additionally, some studies suggest that delayed feedback may benefit certain students depending on their profile (Vanacore et al., 2024). In fact, a systemic review of learning outcomes relating to intelligent tutoring systems found that they may be as effective as individualized human tutoring (Ma et al., 2014).

While it is easy to get carried away amidst the excitement, it is important that we remain grounded and consider the role of ITS within the broader context of curriculum, schools, and teachers towards achieving the desired educational outcomes. Successful ITS at scale, such as ALEKS, ASSISTments, and CognitiveTutor, often aim to complement teachers rather than replace them entirely (Heffernan and Heffernan, 2014; Baker, 2003; Pane et al., 2014). Baker (2016) emphasizes the importance of intelligent humans in leveraging the learning analytics provided by these tutoring systems to better support student learning, or as Koedinger aptly puts it, "essentially freeing the teacher to do the magical things that only great teachers can do" (Koedinger and Aleven, 2016). These ITS platforms provide teachers with learning analytics reports to support their teacher practice. However, the extent to which teachers utilize these data-driven insights remains an open question.

2.2. Teachers' Engagement with ITS and Learning Analytics

Despite the promise of ITS, their effectiveness depends not only on student engagement but also on how teachers integrate them into classroom instruction. A systemic review of barriers to ITS adoption noted how successful ITS at scale such as ALEKS, ASSISTments, and CognitiveTutor generally feature well-developed analytics reports (Nye, 2014). These reports help teachers monitor student progress and provide insights into students' knowledge attainment and gaps in understanding. The review speculated that their success was due in part to these features since they helped to save teachers' time and supported their teaching pedagogies.

Research studies such as Xhakaj et al. (2016) and (2017) have found that teachers who utilize data from such reports are better equipped to support their students' learning. This may be manifested in a variety of ways, ranging from class-level decisions such as lesson planning and selecting areas of remediation, to identification of specific students who may require additional attention. Similarly, Kelly et al. (2013) demonstrated how these reports could be used anonymously in the classroom to facilitate discussion with students, yielding positive student learning outcomes. Holstein et al. (2017), Holstein et al. (2018), and Aleven et al. (2022) also showed that real-time use of these reports helped teachers become more aware of their students' progress and the challenges they encountered with the assigned work. This awareness enabled timely interventions or even encouragements to keep students on-task, since students recognized that teachers were paying attention and monitoring their efforts.

Nonetheless, these studies were largely small-scale in nature, involving either classroom observations or analysis of student outcomes involving a small number of teachers. There are some quantitative research studies that have examined teachers' use of ITS through log data (for example, Vanacore et al. 2021 and Helsabeck et al. 2022) which investigate whether teachers are using ITS as intended. Vanacore et al. (2021) investigated clustering of teachers based on login behavior and its association with student outcomes, while Helsabeck et al. (2022) conducted a randomized controlled trial involving 30 teachers to evaluate fidelity of implementation in a technology-based intervention. However, these studies remain either small scale or observational studies utilizing association analysis which limits further causal inference in understanding the impact on student outcomes when teachers use ITS as intended. In fact, a systemic review of research on teachers' use of analytics from ITS found that about half of the research cases did not include evidence of the claims of the benefits to teachers and students, with observations and self-reporting forming the majority of evidence if any (Ley et al., 2023).

2.3. Randomized Encouragement Design

Randomized Controlled Trials (RCTs) are widely regardly regarded as the gold standard for causal inference in education and social behaviour science research (Xiao et al., 2024; Ginsburg and Smith, 2016). However, in many educational settings, RCTs are often neither feasible nor practical.

For instance, a traditional RCT may randomly assign teachers or students to a particular treatment or intervention that is believed to be beneficial, such as students participating in a preschool programme or teachers receiving professional training. However, enforcing full compliance may be unrealistic since teachers and students retain agency over their behaviour and may choose whether to engage in the treatment, regardless of assignment. Even if we could enforce strict compliance through external means to isolate the causal impact of such practice, such an approach may be impractical or unethical due to potential disruptions to teachers and students.

When a traditional RCT is not feasible or ethical, researchers often use alternatives such as observational studies or quasi-experiments (West et al., 2008). However, these methods can face challenges in ensuring reliable causal conclusions. The presence of unobserved confounders in observational studies may lead to biased estimates while strong assumptions are required for quasi-experiments which may not always hold, making the results sensitive to model specification.

The randomized encouragement design (RED) addresses these limitations by randomly assigning an encouragement to participate in the intended treatment of interest while keeping participation voluntary (West et al., 2008; Bradlow, 1998). Nonetheless, it requires its own set of assumptions and is not without its own limitations. RED belongs to a broader class of causal inference methods known as instrumental variable (IV) analyses. Unlike RCTs, which allow researchers to estimate the *average treatment effect (ATE)* for the entire population (assuming full compliance is possible), RED enables researchers to estimate the *local average treatment effect (LATE)*—that is, the effect of the treatment among compliers, i.e., those who can be successfully encouraged to participate in the treatment.

This design does not provide treatment effect estimates for "always-takers" (those who would participate regardless of encouragement) or “never-takers” (those who would not participate even if encouraged), and therefore the LATE may not generalize to the broader population (Angrist et al., 1996). Nonetheless, in many policy and research settings, interest often lies in understanding the treatment effects among individuals whose behavior can be influenced through an intervention, which corresponds to the *complier group* in RED and IV analyses. For instance, Schochet and Chiang (2011) states that LATE is “often of policy interest, because it pertains to intervention effects for students who receive a meaningful dose of treatment services”.

Although RED is less common than RCTs, it has been used in prior education research to evaluate interventions where full treatment compliance cannot be enforced and the assumptions are reasonably defensible (Angrist and Lavy, 2009; Paloyo et al., 2016; Keller and Szakál, 2021). This study employs RED to investigate the causal impact of teachers’ use of ITS reports on student outcomes. The detailed implementation of RED in our experiment, including statistical assumptions and estimation strategy, is outlined in the Methodology section.

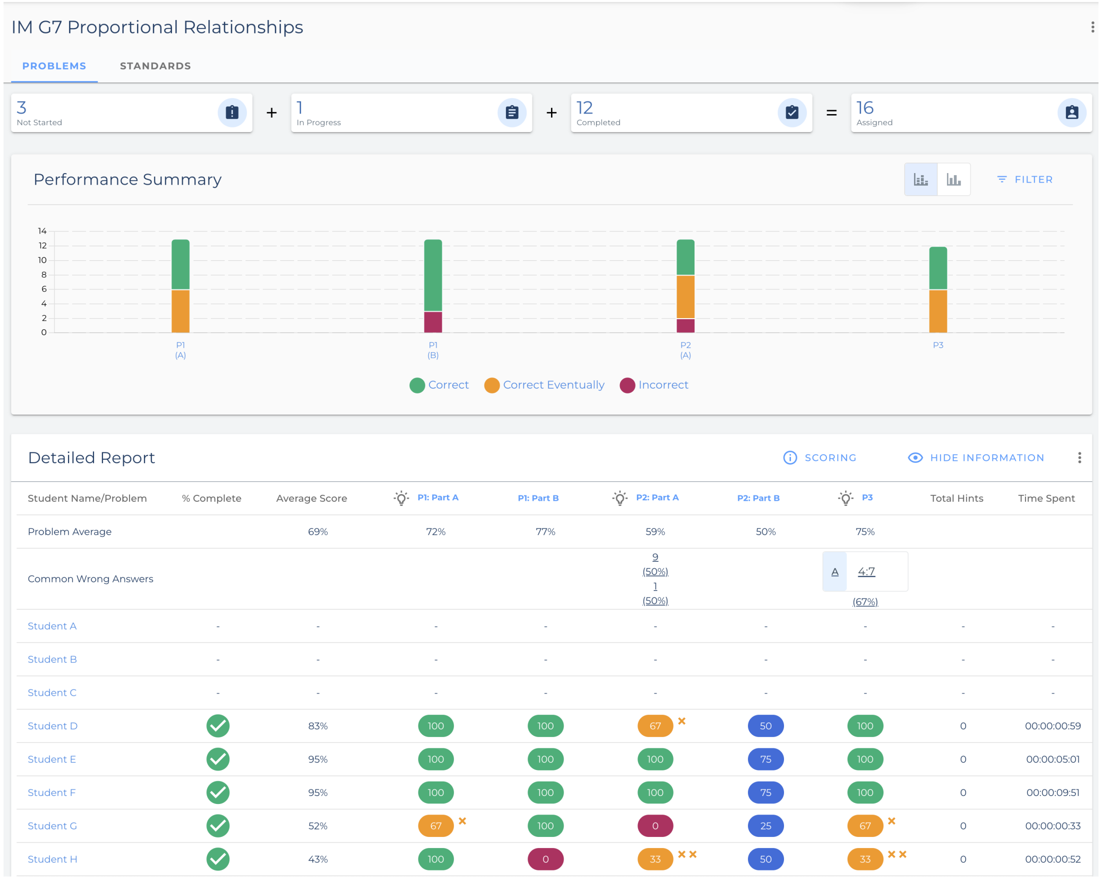
3. Methodology

3.1. ASSISTments and Assignment Reports

To investigate impact of teachers’ use of the assignment report on student outcomes more quantitatively, this study utilizes data from ASSISTments (Heffernan and Heffernan, 2014), an online math platform designed to support teachers by providing mathematics content and problems from a variety of open educational resources (such as Illustrative Mathematics and Engage New York). The platform enables teachers to create assignments for their students based on problem sets from their designated curriculums or curate problems as needed. They are also able to assign “skillbuilders”, a set of problems curated by ASSISTments focused on a specific mathematics skill where students need to answer three consecutive problems correctly to complete the assignment.

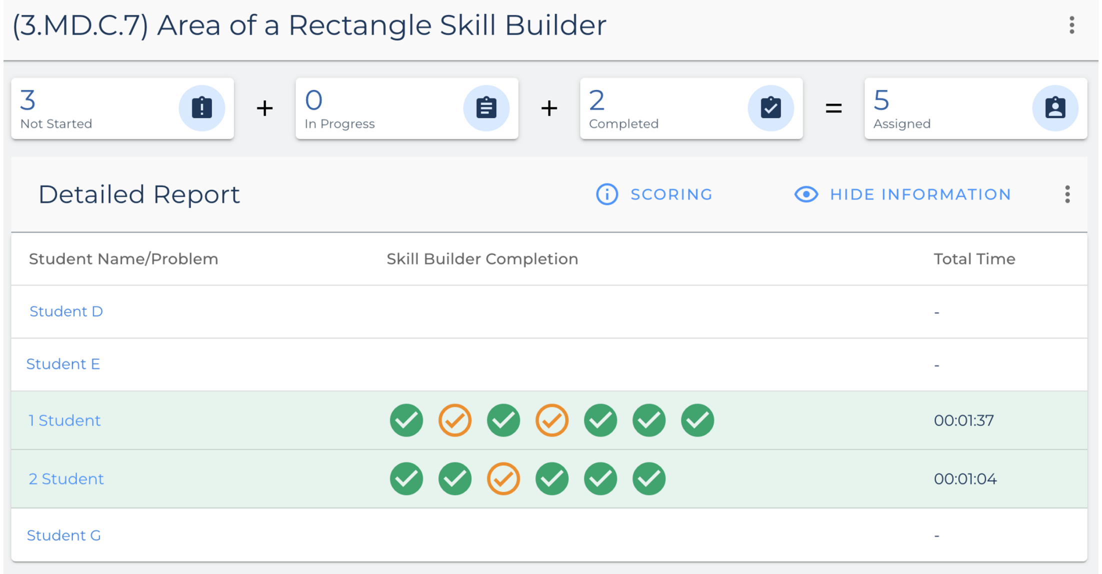
For both problem set and skillbuilder assignments, teachers are able to monitor student progress live in the assignment report (see Figures 1 and 2). These reports allow teachers to track which students have started or completed the assignment and to identify any who are struggling through color-coded indicators: (i) green (solved on first attempt without hints); (ii) orange (solved eventually with hints); and (iii) red (needed to see the answer to progress). While skillbuilder assignments focus on a specific mathematics skill, problem set assignments may span multiple mathematics skills. As such, the problem set assignment report further helps teachers identify the specific problems which posed difficulties for the students, facilitating targeted remediation or classroom discussion if required. For both problem sets and skillbuilders, teachers may drill into the student details report (see Figure 3) to get deeper insights into individual students’ work on the assignment.

Teachers who use the assignment report gain valuable insights into student performance, which can inform their instructional decisions. However, it remains unclear how frequently teachers utilize these reports and whether their use impacts student learning. This study examines whether encouraging teachers to view the report increases student participation and improves performance.



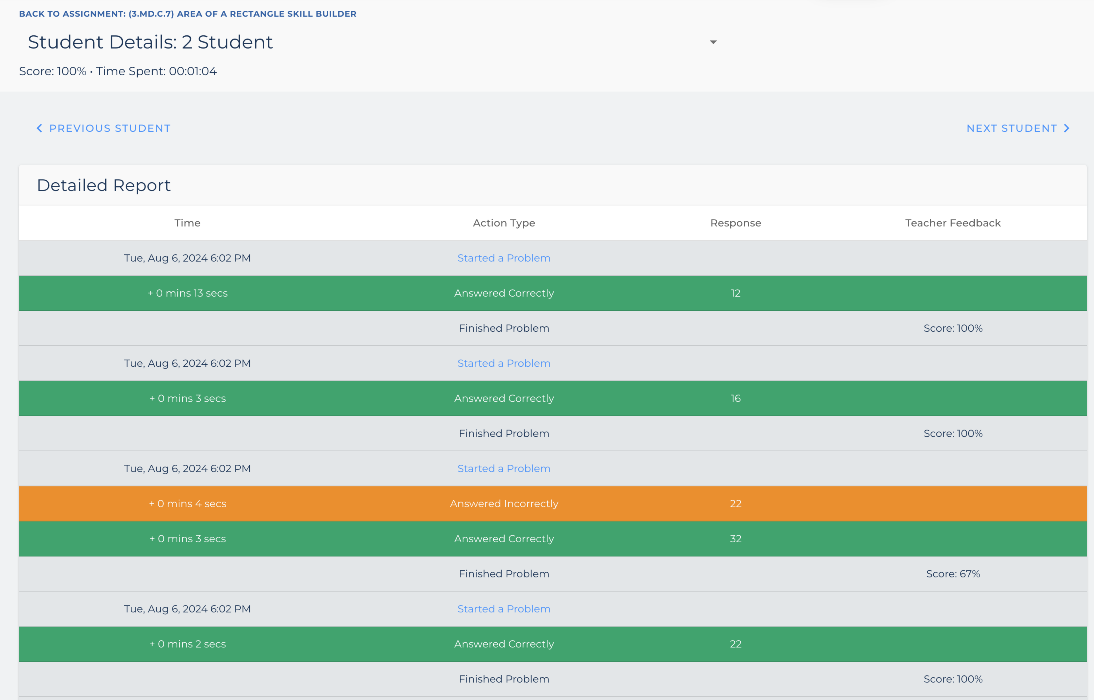
Student Name/Problem	% Complete	Average Score	P1: Part A	P1: Part B	P2: Part A	P2: Part B	P3	Total Hints	Time Spent
Problem Average		69%	72%	77%	59%	50%	75%		
Common Wrong Answers					9 (50%) 1 (50%)		A 4:7 (67%)		
Student A	-	-	-	-	-	-	-	-	-
Student B	-	-	-	-	-	-	-	-	-
Student C	-	-	-	-	-	-	-	-	-
Student D	✓	83%	100	100	67 ×	50	100	0	00:00:00:59
Student E	✓	95%	100	100	100	75	100	0	00:00:05:01
Student F	✓	95%	100	100	100	75	100	0	00:00:09:51
Student G	✓	52%	67 ×	100	0	25	67 ×	0	00:00:00:33
Student H	✓	43%	100	0	33 × ×	50	33 × ×	0	00:00:00:52

Figure 1: Assignment report for problem set



Student Name/Problem	Skill Builder Completion	Total Time
Student D		-
Student E		-
1 Student	✓ ✓ ✓ ✓ ✓ ✓ ✓	00:01:37
2 Student	✓ ✓ ✓ ✓ ✓ ✓	00:01:04
Student G		-

Figure 2: Assignment report for skillbuilder


BACK TO ASSIGNMENT: (3.MD.C.7) AREA OF A RECTANGLE SKILL BUILDER
‹ PREVIOUS STUDENT
NEXT STUDENT ›

Student Details: 2 Student

Score: 100% · Time Spent: 00:01:04

Detailed Report

Time	Action Type	Response	Teacher Feedback
Tue, Aug 6, 2024 6:02 PM	Started a Problem		
+ 0 mins 13 secs	Answered Correctly	12	
	Finished Problem		Score: 100%
Tue, Aug 6, 2024 6:02 PM	Started a Problem		
+ 0 mins 3 secs	Answered Correctly	16	
	Finished Problem		Score: 100%
Tue, Aug 6, 2024 6:02 PM	Started a Problem		
+ 0 mins 4 secs	Answered Incorrectly	22	
+ 0 mins 3 secs	Answered Correctly	32	
	Finished Problem		Score: 67%
Tue, Aug 6, 2024 6:02 PM	Started a Problem		
+ 0 mins 2 secs	Answered Correctly	22	
	Finished Problem		Score: 100%

Figure 3: Student details report

3.2. Experiment Setup and implementation

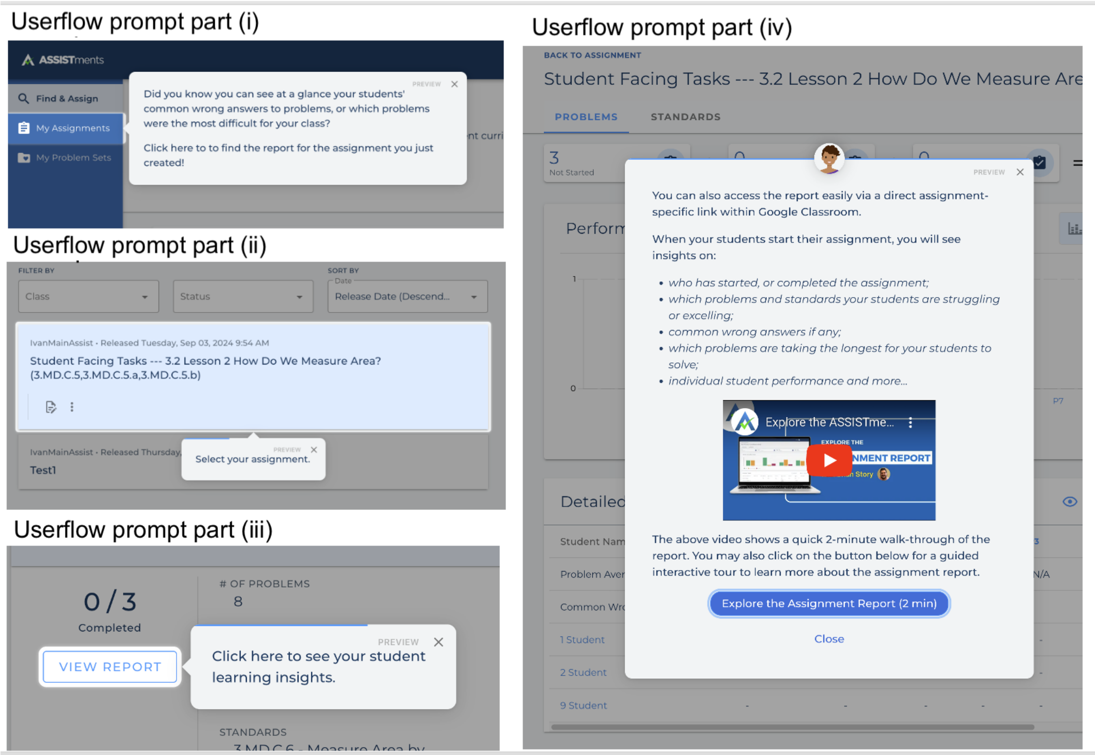
ASSISTments uses Userflow (Userflow, 2024) for onboarding and guided tours. By tapping on Userflow's A/B testing feature (Userflow, 2023), teachers were randomly assigned with equal probability into two groups: control group (group A) or encouragement group (group B), when they logged on to the platform after the start of the experiment. When teachers in group B first created an assignment after the experiment was turned on, they received an encouragement prompt at the end of the assignment creation sequence, urging them about the benefits of the assignment report and to review them, as shown in Figure 4. In contrast, teachers in group A received no such encouragement prompts (i.e., business as usual), with an invisible Userflow launcher capturing these instances.

Previous Userflow analysis within ASSISTments indicate that approximately 10% to 15% of teachers may not be detected by Userflow due to school network policies or browser incompatibilities. By capturing assignment creation instances for both the encouragement group (group B) and the control group (group A), the study ensured minimal bias from undetected teachers.

3.3. Experimental Design: Randomized Encouragement Design

While a traditional randomized controlled trial would randomly assign teachers to either use or not use the assignment report, such an approach is infeasible in practice, as teachers retain full autonomy over their actions on the ASSISTments platform. Researchers cannot enforce whether a teacher views the report, making compliance with a treatment assignment difficult to guarantee.

This presents a key challenge for causal inference: the decision to view the assignment report is likely endogenous. That is, report usage may be influenced by unobserved factors—such as



Userflow prompt part (i)

Did you know you can see at a glance your students' common wrong answers to problems, or which problems were the most difficult for your class?

Click here to to find the report for the assignment you just created!

Userflow prompt part (ii)

IvanMainAssist • Released Tuesday, Sep 03, 2024 9:54 AM

Student Facing Tasks --- 3.2 Lesson 2 How Do We Measure Area? (3.MD.C.5,3.MD.C.5.a,3.MD.C.5.b)

IvanMainAssist • Released Thursday,

Test1

Select your assignment.

Userflow prompt part (iii)

Click here to see your student learning insights.

Userflow prompt part (iv)

Student Facing Tasks --- 3.2 Lesson 2 How Do We Measure Are

You can also access the report easily via a direct assignment-specific link within Google Classroom.

When your students start their assignment, you will see insights on:

- *who has started, or completed the assignment;*
- *which problems and standards your students are struggling or excelling;*
- *common wrong answers if any;*
- *which problems are taking the longest for your students to solve;*
- *individual student performance and more...*

The above video shows a quick 2-minute walk-through of the report. You may also click on the button below for a guided interactive tour to learn more about the assignment report.

Figure 4: Screenshots of Encouragement Prompts

teacher motivation, instructional quality, or access to resources—that also directly affect student outcomes. For example, teachers in well-resourced schools may have smaller class sizes or more instructional support, allowing them both to use data reports more frequently and to provide more targeted instruction. Similarly, more experienced or dedicated teachers may be both more inclined to view reports and more effective at improving student learning. In such cases, any correlation between report usage and student outcomes may simply reflect underlying teacher or school characteristics, rather than the causal effect of viewing the report.

To address this endogeneity, we adopt a randomized encouragement design, where teachers are randomly assigned to receive an encouragement prompt to view the assignment report but ultimately retain the freedom to decide whether to do so. This design introduces variation in report usage that is not driven by teacher characteristics or other confounding factors, by using the randomized encouragement as an instrumental variable—a variable that influences whether teachers view the report but is otherwise unrelated to factors that also affect student outcomes.

Importantly, the encouragement itself has no direct effect on students: it is only visible to teachers and does not include any instructional content or student-specific information. As such, its influence on student outcomes operates solely through its effect on whether teachers view the assignment report. The logic of this design is illustrated in Figure 5.

However, since not all teachers comply with their assigned condition—some may ignore the encouragement, while others may view the report even without being prompted—the actual treatment (report usage) is not randomly assigned, even though it originates from the randomized encouragement. As such, we rely on an instrumental variables approach to estimate the causal effect of viewing the assignment report on student outcomes. Since the instrument (encourage-

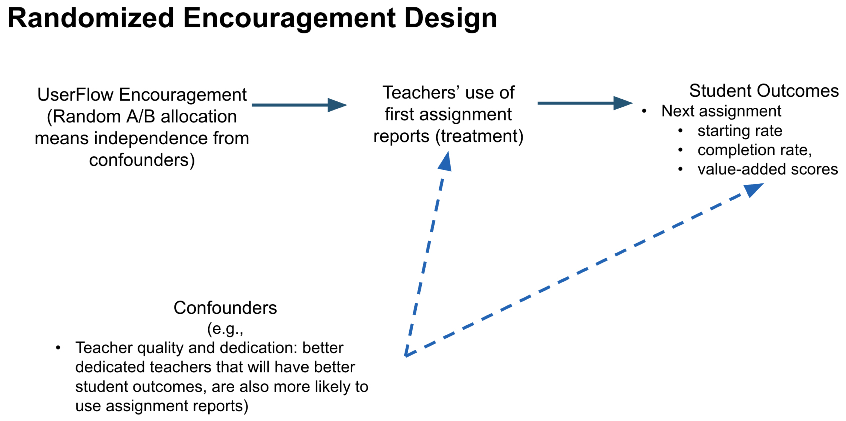

Randomized Encouragement Design
UserFlow Encouragement (Random A/B allocation means independence from confounders)
Teachers' use of first assignment reports (treatment)
Student Outcomes
Next assignment
starting rate
completion rate,
value-added scores
Confounders (e.g.,
Teacher quality and dedication: better dedicated teachers that will have better student outcomes, are also more likely to use assignment reports)

Figure 5: A causal diagram that shows the Userflow encouragement as an instrumental variable. *Note: Dashed lines indicate potentially unobserved relationships.*

ment) in RED is already randomized, the validity of this approach depends on the following three assumptions (Lousdal, 2018):

1. **Relevance:** The encouragement works, i.e., there must be a positive effect of the encouragement in increasing the probability that the teacher views the report.

    - This assumption is empirically testable and is addressed in Research Question 1.

2. **Exclusion Restriction:** The encouragement affects student outcomes only through its influence on whether the teacher views the assignment report.

    - While we cannot empirically test the exclusion restriction, we argue that this assumption is reasonable, as the encouragement is only visible to teachers within the ASSISTments platform and is not accessible to students. Hence, it can only affect students by influencing teacher behavior—most plausibly through report viewing, which is the intended target of the encouragement.
    - We note that any unintended behavioral effects triggered by the encouragement message would need to systematically affect student outcomes in a way not mediated by report viewing—which is unlikely given the limited nature of the message.
    - In addition, teachers do not receive additional information or pedagogical support via the encouragement message itself, and the message contains no instructional content or performance information that could directly impact teaching practices. While we cannot completely rule out minor behavioral changes triggered by the encouragement message itself, its shortness would likely minimize such effects. Thus, any direct influence of encouragement on student outcomes outside of report viewing is likely negligible.

3. **Monotonicity:** There are no defiers, i.e., there are no teachers who would view the report only if *not* encouraged and refuse to view it if encouraged.

    - We consider this assumption plausible, as it would require teachers to behave in direct opposition to their natural inclination. For instance, a defier would be someone who deliberately avoids the report after receiving the encouragement, even though they would have viewed it otherwise.

- Given that all teachers have access to the assignment report by default, and that teachers in the control group are unaware of the encouragement condition, any indirect exposure—such as hearing about the prompt from encouraged peers—is likely to be rare and any resulting influence would likely be minimal and not systematically violate the monotonicity assumption.

In the next subsection, we describe how we estimate the treatment effect among compliers—those who can be successfully encouraged to participate in the treatment—using the two-stage least squares (2SLS) approach.

3.4. Estimation of Treatment Effects in RED

To estimate the causal effect of teachers viewing the assignment report on student outcomes, we employ a two-stage least squares (2SLS) approach, using random assignment to encouragement as an instrumental variable. This is necessary because actual report usage is endogenous—that is, teachers decide whether to view and engage with the report, but this decision may be correlated with unobserved factors such as teacher motivation or instructional quality, which also affect student outcomes.

Two-Stage Least Squares (2SLS). The 2SLS estimation proceeds in two stages (Martens et al., 2006). For a just-identified model with a single instrument and one endogenous regressor, we first model in the first-stage regression whether teacher i viewed the assignment report (D_i) as a function of their encouragement assignment (Z_i):

$$D_i = \gamma + \delta Z_i + \varepsilon_i \tag{1}$$

Here, Z_i is a binary instrumental variable indicating whether teacher i was assigned to the encouragement group, and D_i is a binary variable indicating whether the teacher viewed the assignment report. The coefficient δ captures the strength of the instrument—that is, how much the encouragement increases the likelihood of report usage. The intercept γ represents the baseline probability of report viewing among teachers in the control group, while ε_i captures any remaining unobserved factors affecting the teacher's likelihood of viewing the report.

This first-stage regression isolates the variation in D_i that is solely due to the randomized encouragement. Thus, the resulting fitted values $\hat{D}_i$ represent the predicted probability of report usage based solely on the instrument, filtering out any endogenous variation due to unobserved teacher factors such as teacher motivation.

In the second-stage regression, we estimate the outcome of interest Y_i using the fitted values from the first stage for D_i:

$$Y_i = \alpha + \beta \hat{D}_i + \eta_i \tag{2}$$

The dependent variable Y_i may be a student-level continuous outcome (e.g., value-added score), or a binary indicator (e.g., whether a student started or completed the assignment). When aggregated at the teacher level, Y_i can also represent continuous outcomes of proportions (e.g., proportion of students completing the assignment). The coefficient β represents the local average treatment effect—the causal effect of report usage on student outcomes for the subgroup of teachers whose behavior was influenced by the encouragement (i.e., compliers), while α captures the expected outcome for the control group. The error term η_i reflects the unobserved factors affecting outcome Y_i that are not explained by the teacher's predicted report usage $\hat{D}_i$.

For clarity, although 2SLS is conceptually a two-stage procedure, the estimation is not typically performed in two completely separate steps. Instead, matrix-based methods are used to implicitly combine both stages into one structural equation to be solved. This approach ensures that the standard errors in the second stage appropriately account for the uncertainty from the first-stage estimation.

Use of Linear Probability Model (LPM). We use a linear probability model (LPM) for both the first and second stage when the dependent variable is binary. The results from this model are similar to those from logistic or probit regression analyses when the estimated probabilities are between 0.2 and 0.8 (Martens et al., 2006).

This avoids complications associated with applying logistic regression in IV settings, which is known to yield inconsistent estimates unless strong distributional assumptions are met (Angrist and Pischke, 2008; Angrist, 2001). Although the LPM can produce predicted probabilities outside the [0,1] range and may exhibit heteroskedasticity, it provides a consistent and interpretable estimate of average causal effects, particularly LATE. Robust standard errors are used throughout to account for heteroskedasticity.

3.4.1. Estimation in the Presence of a Weak Instrument.

The encouragement prompt in this study was designed to be unobtrusive and minimally disruptive to teachers' existing workflows on ASSISTments. This design choice helps reduce the risk of violating the exclusion restriction assumption, i.e., it limits the likelihood that the encouragement influences student outcomes through channels other than report usage (e.g., by prompting unrelated changes in teaching practices).

However, this conservative implementation also increases the risk that the instrument is weak—that the encouragement may have only a modest effect on whether teachers actually view the assignment report. A weak instrument can lead to biased and imprecise estimates of the treatment effect, and can compromise the validity of standard 2SLS inference (Bound et al., 1995). A commonly used rule-of-thumb is that the first-stage F-statistic should exceed 10 (Staiger and Stock, 1997) to minimize issues arising from a weak instrument.

Since weak instrumentation is a valid concern, we account for this possibility by employing two approaches to minimize the risk of biased estimation. Specifically, in the event that the encouragement leads to a statistically significant but modest increase in the proportion of teachers who viewed the assignment report, we adopt the following two methods of inference.

- **Anderson-Rubin (AR) Test:** Unlike 2SLS, the AR test does not yield a point estimate of the treatment effect but instead provides inference on whether the effect is statistically different from zero. The AR test constructs confidence intervals for the treatment effect that remain valid even when the instrument is weak. Unlike conventional t-tests, the AR test does not rely on strong first-stage relationships and remains robust under weak identification. Furthermore, Keane and Neal (2024) and Andrews et al. (2019) recommend using the Anderson-Rubin (AR) test over the traditional two-tailed t-test in two-stage least squares (2SLS), even with strong instruments.

  In the just-identified case with a single instrument and endogenous regressor such as ours, the AR test of $H_0 : \beta = 0$ can be implemented as the t-test from regressing the outcome Y

on the predicted value of the endogenous variable $\hat{D}$ obtained from the first-stage regression, and yields valid inference even when instruments are weak (Keane and Neal, 2024; Andrews et al., 2019).

- **Fuller (1) Minimum-Bias Estimator:** In just-identified IV models such as ours, the 2SLS estimator is approximately unbiased, even in the presence of weak instruments (Angrist and Pischke, 2008). However, it can still be highly imprecise in finite samples due to the heavy-tailed nature of its sampling distribution.

  The Limited Information Maximum Likelihood (LIML) estimator addresses these concerns by estimating the same structural equation as 2SLS ($Y = \alpha + \beta D + \eta$), but is based on optimizing a likelihood-like function that often yields more stable estimates in small samples (Greene, 2018). Hayashi (2000) describes LIML as the "maximum likelihood counterpart of 2SLS" while Burgess et al. (2017) notes that it often tends to provide similar causal estimates in models such as ours. Both 2SLS and LIML belong to a broader class of estimators known as k-class estimators.

  From Shi (2024), the k-class estimator has the general form:

$$\hat{\beta}^{k-\text{class}} = \frac{D'(I_n - kM_Z)Y}{D'(I_n - kM_Z)D}, \tag{3}$$

  where $M_Z = I_n - Z(Z'Z)^{-1}Z'$, and I_n is the $n \times n$ identity matrix with n representing the number of observations (in this case, teachers included in the analytic sample). Z is the $n \times 1$ binary instrument vector indicating assignment to the encouragement group or control group ; D is the binary treatment vector (whether the teacher viewed the report); and Y is the outcome vector.

  The LIML estimator corresponds to the smallest root $\tilde{k}$ of the equation:

$$\det\left([D,Y]'[D,Y] - k[D,Y]'M_Z[D,Y]\right) = 0. \tag{4}$$

  To reduce bias in finite samples, the Fuller(1) estimator adjusts the LIML eigenvalue:

$$k_{\text{Fuller}(1)} = \tilde{k} - \frac{1}{n-1} \tag{5}$$

  This corrected k is then plugged into the k-class estimator formula above to compute the Fuller(1) estimate. The Fuller(1) estimator has been shown to perform well in simulations, particularly under weak instrumentation and small samples (Andrews and Armstrong, 2017; Hahn et al., 2004).

Importantly, Angrist and Kolesár (2024) note that in models with a single instrument and a single endogenous variable—such as our design—standard inference methods, including 2SLS, tend to perform reasonably well. Nonetheless, the use of the AR test and the Fuller (1) estimator provide added robustness in the presence of potentially weak instrumentation.

3.4.2. Implementation Using ivmodel in R

All instrumental variable analyses were conducted in R version 4.3.3 using ivmodel version 1.9.1. ivmodel supports a suite of estimation and inference methods for linear IV models (Kang et al., 2021). Specifically, we estimated the treatment effect using two-stage least squares (2SLS, or TSLS in the package), the Anderson–Rubin (AR) test, and the Fuller(1) estimator by implementing the ivmodel() function. For clarity, Fuller(1) is the default implementation in ivmodel function, although other *k* values of Fuller(*k*) can be found using the package.

Where applicable, we set the option for heteroSE, (i.e., heteroskedastic-robust standard errors) to TRUE, and specified relevant groups (i.e., students in the same class working on the same assignment) using clusterID supported in the ivmodel function. These features enable estimation of valid confidence intervals and test statistics under both heteroskedastic and clustered error structures, which is essential given the hierarchical structure of our data and the binary nature of some of the outcome variables.

3.5. Data Collection

As explained in section 3.4.1, in randomized encouragement design experiments one of the biggest risks is due to the "weak instrument problem" (Andrews et al., 2019), where the instrument (i.e., the encouragement in this case) is not strong enough to differentiate the encouragement group and the control group. Since the encouragement on Userflow may not be effective to teachers who are already familiar with the assignment report, the target group of teachers in the experiment are teachers new to ASSISTments where the encouragement may be more effective. This group is likely to include more "compliers". For clarity, "teachers new to ASSISTments" refers specifically to teachers new to the platform and not "new teachers", as ASSISTments does not collect any data on teaching experience.

To investigate the causal impact of teachers viewing the assignment report, the analysis would focus on the first two assignments of teachers new to ASSISTments after the experiment begins, where we look at the impact of the teacher viewing the assignment report for the first assignment on the starting rate, completion rate, and student performance in the second assignment. To evaluate student performance, we would use "value-added scores" based on how these students performed compared to how other students have done the same problems in the past since July 2022.

Figure 6 shows the flow of participants through the study, starting from the initial pool of teachers who logged into ASSISTments during the experiment period which began in late August 2024 and ended in early December 2024. Of the 2,808 teachers who logged in (1,368 in the control group and 1,440 in the encouragement group), the target population was narrowed to the 589 teachers who were new to ASSISTments and had created at least one assignment before November 10th 2024—276 in the control group and 313 in the encouragement group. While the size of teachers in the encouragement group is slightly larger than the size of teachers in the control group teachers, we note that the proportion of teachers that are in the control group are not statistically different from half in both the initial pool and the target population, as expected from the randomization.

To isolate the causal impact of teachers viewing an assignment report on the next assignment, we considered two analytic samples:

- Analytic sample 1 (330 teachers: 156 control, 174 encouragement): to include only teachers who created at least two assignments that are not ASSISTments orientation problem

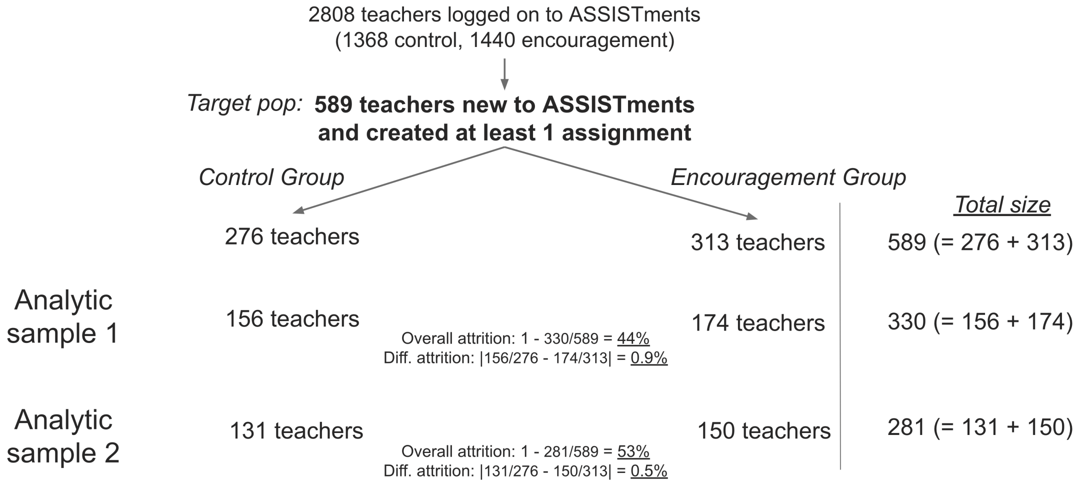

2808 teachers logged on to ASSISTments
(1368 control, 1440 encouragement)
Target pop: 589 teachers new to ASSISTments and created at least 1 assignment
Control Group
Encouragement Group
Total size
276 teachers
313 teachers
589 (= 276 + 313)
Analytic sample 1
156 teachers
Overall attrition: 1 - 330/589 = 44%
Diff. attrition: |156/276 - 174/313| = 0.9%
174 teachers
330 (= 156 + 174)
Analytic sample 2
131 teachers
Overall attrition: 1 - 281/589 = 53%
Diff. attrition: |131/276 - 150/313| = 0.5%
150 teachers
281 (= 131 + 150)

Figure 6: Flow diagram of teacher sample selection and attrition

sets since these problem sets did not require teachers to follow up on the report

  - This analytic sample is used for RQ1, RQ2a1, RQ2a2, RQ2b1 and RQ2b2.
  - It allows us to analyze the causal impact of teachers viewing the assignment report for the first assignment on the proportion of students that started and completed the second assignment

- Analytic sample 2 (281 teachers: 131 control, 150 encouragement): to include only teachers in the first analytic sample who used non-open response problem sets for the second assignment

  - This analytic sample is only used for RQ2c and allows us to analyse "value-added scores" to avoid bias resulting from teachers' grading or from teachers' neglecting to grade the open-response items, as well as dropping skillbuilders which needed students to get three consecutive problems right to complete the assignment.

This approach resulted in two sets of attrition. As our randomized encouragement design required the estimation of the local average treatment effect from the non-attritors, we assumed that the attrition did not lead to any bias due to differences in the teachers who remained in the analytic sample. Following the recommendations of the What Works Clearinghouse (What Works Clearinghouse, 2022), we compared the overall and differential attrition rates and found both analytic samples to have attrition rates to fall within the cautious boundary for differential attrition (see Figure 7). This suggests that the levels of overall and differential attrition are considered low enough to minimize potential bias. While we do not expect the encouragement to affect the attrition in both analytic samples in any way, the cautious boundary is chosen as an extra safeguard in case the encouragement increased the likelihood of teachers having a second assignment compared to the control group.

We also assessed baseline covariate balance between the encouragement and control groups for both analytic samples. While balance checks are not required under What Works Clearinghouse (WWC) standards when attrition is considered low (What Works Clearinghouse, 2022), we conducted them to strengthen the internal validity of the findings. Across both samples, we observed no substantial imbalances in observed covariates, with all absolute Cox/Hedges' g values below 0.25 (see Tables 1 and 2), suggesting that the observed differences are unlikely to introduce bias.

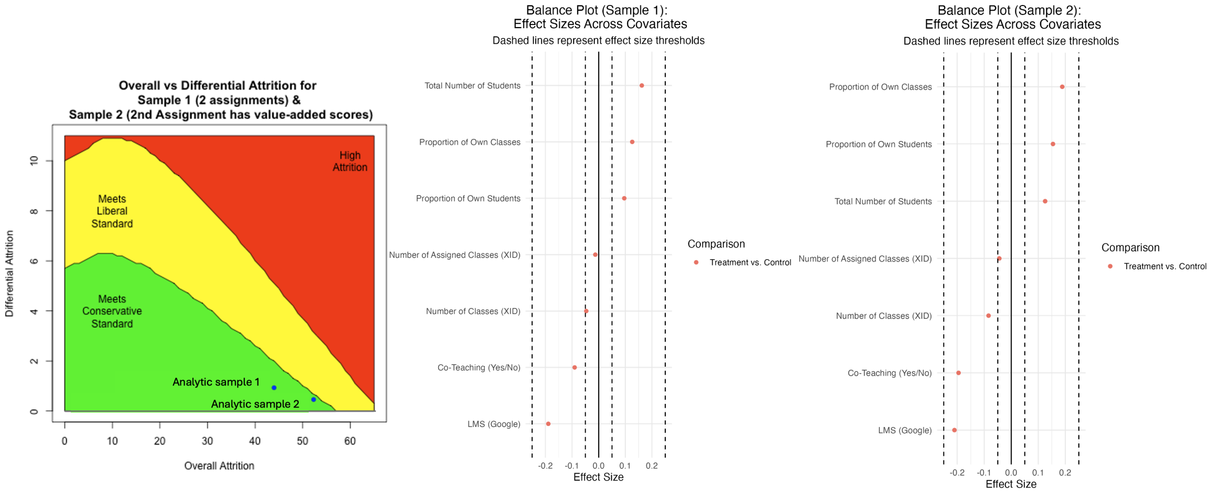


Figure 7: Overall vs Differential Attrition Rates, Covariates Balance for Analytic Samples 1 & 2

Table 1: Covariate Balance Between Encouragement and Control Groups in Analytic Sample 1

Covariate	Mean_Control	Mean_Treatment	Mean_Difference	Cox/Hedges' G
Total Number of Students	46.78	52.93	6.15	0.16
Proportion of Own Classes	0.96	0.98	0.02	0.13
Proportion of Own Students	0.96	0.98	0.02	0.10
Number of Assigned Classes	2.42	2.40	-0.02	-0.01
Number of Classes	2.49	2.42	-0.07	-0.05
Co-Teaching (Yes/No)	0.28	0.25	-0.03	-0.09
LMS (Google)	0.74	0.67	-0.06	-0.19

Note: The table is sorted in descending order of Cox/Hedges' G values.

Table 2: Covariate Balance Between Encouragement and Control Groups in Analytic Sample 2

Covariate	Mean_Control	Mean_Treatment	Mean_Difference	Cox/Hedges' G
Proportion of Own Classes	0.95	0.98	0.03	0.19
Proportion of Own Students	0.95	0.98	0.03	0.15
Total Number of Students	48.60	53.46	4.87	0.13
Number of Assigned Classes	2.47	2.40	-0.07	-0.04
Number of Classes	2.56	2.42	-0.13	-0.08
Co-Teaching (Yes/No)	0.31	0.24	-0.06	-0.20
LMS (Google)	0.73	0.65	-0.07	-0.21

Note: The table is sorted in descending order of Cox/Hedges' G values.

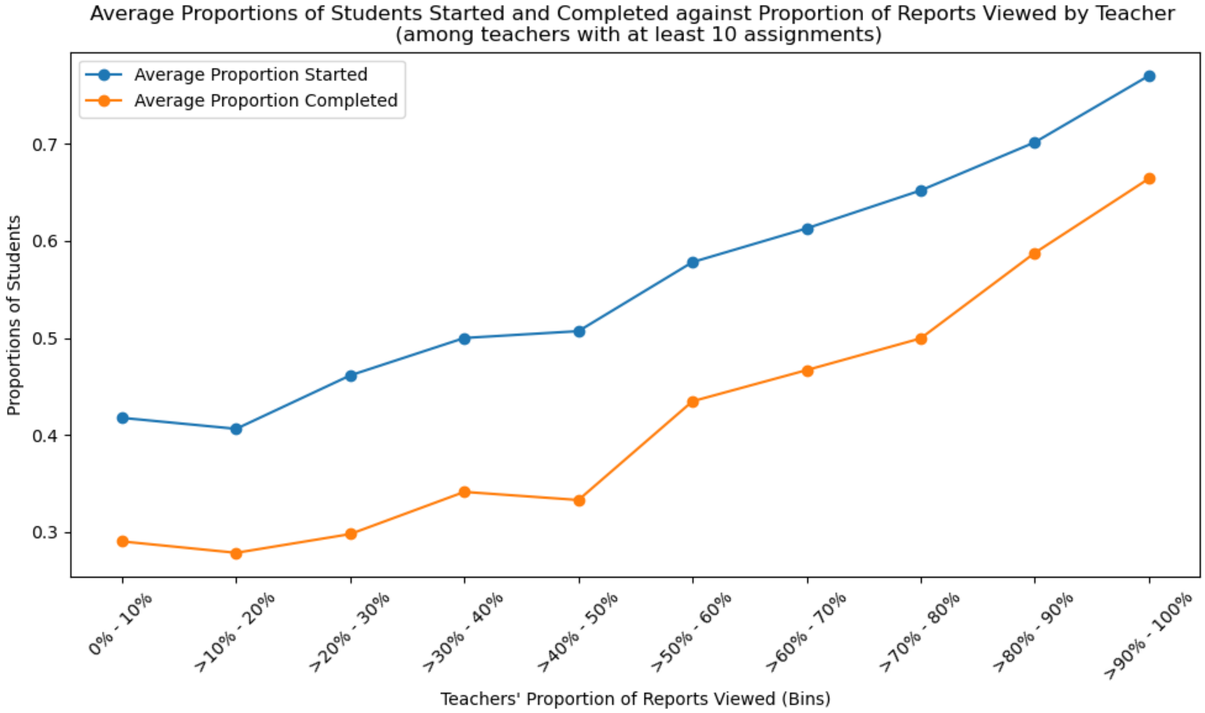


Figure 8: Average assignment starting and completion rates by proportion of assignments viewed

4. Results

4.1. Background data of teachers' use of assignment report in ASSISTments

To better contextualize the causal estimates from the randomized encouragement design, we first examine descriptive patterns in teachers' usage of the assignment report across the ASSISTments platform. Understanding how report usage varies in practice—and how it correlates with student participation—provides important background for interpreting the potential impact of encouraging teachers to engage with these reports. While these correlations do not establish causality, they help illustrate the practical significance and real-world variability of teachers' report usage.

Over the one-year period from 1st July 2023 to 30th June 2024, 2,351 teachers assigned 121,890 assignments. During this period, only 65,877 (54.0%) of assignments had their reports viewed by the teachers that assigned them. A high-level glimpse of the data showed that for assignments where teachers viewed the assignment reports, the average proportions of students who started or completed the same assignments were 74.6% and 62.7% respectively. In contrast, these proportions dropped to 43.0% and 31.1% respectively when teachers did not view the assignment reports.

When we viewed the average starting and completion rates of assignments among teachers with at least 10 assignments, we also found generally higher average proportions of students starting and completing the assignments among teachers with higher proportions of reports viewed (see Figure 8).

The above results suggest that teachers' use of the assignment report is at least associated with higher student participation in their assignments. Nonetheless, we are unable to exclude the possibility that mediating factors, such as teacher quality and dedication, may explain both teachers' use of the assignment report and higher student participation.

To address these endogeneity concerns and move beyond correlational insights, we begin by examining whether the encouragement prompt significantly increased teachers' probability of viewing the assignment report (RQ1). We then estimate causal effects using the randomized encouragement design, allowing us to isolate the impact of report usage on student outcomes (RQ2), specifically for the subgroup of teachers whose behavior was influenced by the encouragement.

4.2. Results of encouragement (RQ1)

We present the results from the instrumental variable analysis in the next two subsections. We begin with the first-stage regression in this subsection, which evaluates the strength of the encouragement in inducing teachers to view the assignment report, as addressed in RQ1. We then proceed to the second-stage results in the next subsection using the instrumental variable approach to isolate the causal effect of teachers viewing the assignment report on student outcomes in the subsequent assignment, as investigated in RQ2.

Table 3 shows the first-stage regression results which summarize the effectiveness of the Userflow encouragement on inducing teachers to view the first assignment report. Compared to 62.8% of teachers in the control group viewing the report for the first assignment, 75.3% of teachers in the encouragement group viewed the report, an increase of 12.5 percentage points which was statistically significant. This suggests that teachers new to ASSISTments can be encouraged to view their assignment reports.

Table 3: First-stage regression results where dependent variable is whether the teacher viewed the first assignment report

Coefficients	**Estimate**	**Std. Error**	**t-value**	**Pr($>|t|$)**	**Significance**
(Intercept)	0.628	0.037	17.131	<2e-16	***
encouragement	0.125	0.051	2.469	0.0141	*

Notes: Significance codes: *** (p<0.001), ** (p<0.01), * (p<0.05), . ($p < 0.1$)
F-statistic: 6.094 on 1 and 328 DF, p-value: 0.0141

4.3. Results of Instrumental Variable Analysis (RQ2)

As shown in Table 3, the first stage F-statistic is 6.09, which falls below the commonly used rule-of-thumb threshold of $F > 10$ proposed by Staiger and Stock (1997) to determine weak instruments. According to Stock-Yogo critical values (Stock and Yogo, 2005), using two-stage least squares (2SLS) inference in this context would be at risk of 20% to 25% maximal size distortion. As highlighted in section 3.4.1, we also employ the following two methods for inference:

- Anderson and Rubin (AR) test: The AR test is robust to weak instrument, providing valid inferences even when the first-stage relationship is weak. Furthermore, Keane and Neal (2024) and Andrews et al. (2019) recommend using the Anderson-Rubin (AR) test over the traditional two-tailed t-test in two-stage least squares (2SLS), even with strong instruments.

- Fuller(1) minimum-bias estimator: While the AR test allows us to test for the presence of an effect, a point estimate is not readily available. The Fuller(1) estimator provides point estimates and has been shown to perform well in simulations (Andrews and Armstrong, 2017). Nonetheless, Fuller(1) cannot fully eliminate issues associated with weak instruments and can still lead to imprecise causal estimates. Thus, we interpret the point estimates obtained with caution.

While the use of stage 1 encouragement as a weak instrument raises concerns, applying the AR test and the Fuller(1) estimator helps mitigate some of these issues. Moreover, Angrist and Kolesár (2024) suggest that for just-identified single-variable instrument settings—like ours, with one instrument (encouragement) and one endogenous variable (viewing the report)—standard inference methods are generally reliable.

4.3.1. Causal impact on students starting the second assignment (RQ2a1, RQ2a2)

From Tables 4 and 5, we observe that teachers viewing the first assignment report has a positive impact on both the proportion of students starting the second assignment and on the probability that a student would start the second assignment. However, these impacts are not statistically significant.

This lack of statistical significance may stem from other coexisting teaching practices among the teachers using ASSISTments. For example, teachers who initiate assignments with students during class time are likely to achieve high starting rates regardless of whether they reviewed the assignment report. Such practices could mask the measurable effect of report viewing on students starting the assignment.

Table 4: Second-stage results where dependent variable is the proportion of students who started the second assignment.

Anderson-Rubin Test	**2SLS / Fuller(1) t-test**					
under F-distribution	**Method**	**Estimate**	**Std. Error**	**t-value**	**Pr($>|t|$)**	**Significance**
$F = 2.442$, p-value = 0.119	2SLS	0.508	0.335	1.517	0.130	
95% CI: [-0.185, 2.235]	Fuller(1)	0.467	0.283	1.649	0.100	
Notes: Significance codes: *** (p<0.001), ** (p<0.01), * (p<0.05), . (p < 0.1)						

Table 5: Second-stage results where the dependent variable is a binary outcome indicating whether a student started the second assignment.

Anderson-Rubin Test	**2SLS / Fuller(1) t-test**					
under F-distribution	**Method**	**Estimate**	**Std. Error**	**t-value**	**Pr($>|t|$)**	**Significance**
$F = 0.784$, p-value = 0.376	2SLS	0.104	0.116	0.896	0.370	
95% CI: [-0.133, 0.332]	Fuller(1)	0.105	0.115	0.916	0.360	
Notes: Significance codes: *** (p<0.001), ** (p<0.01), * (p<0.05), . (p < 0.1)						

4.3.2. Causal impact on students completing the second assignment (RQ2b1, RQ2b2)

From Table 6, the Anderson-Rubin (AR) test (which is robust to weak instrument) indicates that teachers viewing the first assignment report had a statistically significant positive impact on the proportion of students completing the second assignment, on average, among teachers who were successfully encouraged to view the report. As explained in section 3.4.1, the AR test does not provide a point estimate of the magnitude of the increase in the proportion but yields a confidence interval which has a lower bound of 2 percentage points. While this range is wide—and the upper bound exceeds the logical maximum of 100% due to the linear probability model—it still rules out a null effect and supports a positive impact of at least 2 percentage points increase in completion rates.

We refer to 2SLS and Fuller(1) estimators which suggest the true proportion is much higher than the 2%, as these estimators indicate that report viewing could lead to an increase of close to 60% of students completing the assignment. However, this estimate appears unrealistically high and may be an artifact of the weak instrument used. Despite this, the Anderson-Rubin test provides confidence that the impact is positive and statistically significant.

Table 6: Second-stage results where the dependent variable is the proportion of students who completed the second assignment.

Anderson-Rubin Test	**2SLS / Fuller(1) t-test**					
under F-distribution	**Method**	**Estimate**	**Std. Error**	**t-value**	**Pr($>|t|$)**	**Significance**
$F = 4.092$, p-value = 0.044 *	2SLS	0.635	0.343	1.851	0.065	.
95% CI: [0.022, 2.680]	Fuller(1)	0.577	0.285	2.024	0.044	*
Notes: Significance codes: *** (p<0.001), ** (p<0.01), * (p<0.05), . ($p < 0.1$)						

Interestingly, as shown in Table 7, testing the probability of a student completing the assignment revealed a positive effect, though it was not statistically significant. Combined with the large standard errors in Table 6 and the role of aggregation in teacher-level analyses in extracting treatment effects, this suggests treatment heterogeneity based on the number of students a teacher has, i.e., it suggests smaller treatment effects might be observed among teachers with more students compared to those with fewer students.

Table 7: Second-stage results where the dependent variable is a binary outcome indicating whether a student completed the second assignment.

Anderson-Rubin Test	**2SLS / Fuller(1) t-test**					
under F-distribution	**Method**	**Estimate**	**Std. Error**	**t-value**	**Pr($>|t|$)**	**Significance**
$F = 0.125$, p-value = 0.723	2SLS	0.044	0.123	0.356	0.722	
95% CI: [-0.211, 0.281]	Fuller(1)	0.046	0.121	0.379	0.705	
Notes: Significance codes: *** (p<0.001), ** (p<0.01), * (p<0.05), . ($p < 0.1$)						

To explore this further, we divided teachers into approximately equal thirds based on the number of students they had: fewer than 22 students (113 teachers), 22 to 43 students (107 teachers), and more than 43 students (110 teachers). Table 8 shows that the encouragement effect was minimal for teachers in the top third (teachers with more than 43 students), where approximately three-quarters of teachers viewed the assignment report regardless of their group (control or encouragement). This high baseline viewing rate in both groups in the top third resulted in limited variation in the instrument (encouragement assignment), making it infeasible to reliably estimate treatment effects for compliers in this subgroup using 2SLS, AR, or Fuller(1) estimators.

Table 8: Teacher and Student Engagement in Assignments Based on Student Groups

	Proportion of Teachers who Viewed the First Assignment Report				**Average Proportion of Students who Completed Second Assignment**		
Number of Students	**Control**	**Encouragement**	**Average**		**Control**	**Encouragement**	**Average**
More than 43	75%	77%	76%		54%	49%	51%
22 to 43	70%	84%	78%		44%	53%	49%
Less than 22	48%	63%	55%		29%	45%	36%
Overall	63%	75%	69%		41%	49%	45%

Focusing on the bottom two-thirds (teachers with fewer than 44 students) revealed a stronger impact of encouragement on report viewing. However, the encouragement is still not considered a strong instrument, as the corresponding first-stage F-statistic is only 6.98—below the commonly used rule-of-thumb threshold of $F > 10$ proposed by Staiger and Stock (1997). This value can be readily reproduced by rerunning the first-stage regression shown in Table 3.

Nonetheless, as demonstrated in Tables 9 and 10, both the Anderson-Rubin (AR) test and the Fuller(1) estimator suggest that viewing the first assignment report has a statistically significant and more pronounced positive effect on the proportion of students completing the second assignment among teachers with smaller class sizes (i.e., after excluding those in the top third by student count). The lower bound of the 95% confidence interval from the AR test is 0.245, indicating that the estimated effect of teachers' report viewing is at least a 24.5 percentage point increase in completion rates for this subgroup. In addition, the probability that a student completed the assignment was also statistically significant, indicating that when teachers in smaller classes viewed the assignment report, their students were approximately 60 percentage points more likely to complete the assignment.

Table 9: Second-stage results (for non-large sample) where the dependent variable is the proportion of students who completed the second assignment.

Anderson-Rubin Test	**2SLS / Fuller(1) t-test**					
under F-distribution	**Method**	**Estimate**	**Std. Error**	**t-value**	**Pr($>\|t\|$)**	**Significance**
$F = 7.475$, p-value = 0.007 **	2SLS	0.819	0.373	2.195	0.029	*
95% CI: [0.245, 2.962]	Fuller(1)	0.740	0.309	2.391	0.018	*
Notes: Significance codes: *** (p<0.001), ** (p<0.01), * (p<0.05), . (p < 0.1)						

Table 10: Second-stage results (for non-large sample) where the dependent variable is a binary outcome indicating whether a student completed the second assignment.

Anderson-Rubin Test	**2SLS / Fuller(1) t-test**					
under F-distribution	**Method**	**Estimate**	**Std. Error**	**t-value**	**Pr($>\|t\|$)**	**Significance**
$F = 44.517$,	2SLS	0.635	0.104	6.108	1.10×10^{-09}	***
p-value = 2.845×10^{-11} ***	Fuller(1)	0.632	0.103	6.126	9.82×10^{-10}	***
95% CI: [0.443, 0.856]						
Notes: Significance codes: *** (p<0.001), ** (p<0.01), * (p<0.05), . (p < 0.1)						

4.3.3. Causal impact on value-added scores in the second assignment (RQ2c)

Table 11 shows negligible impact from teachers viewing the first assignment report on students' value-added scores in the second assignment. Since value-added scores compare students' performance to that of peers who completed the same problems in the past, significant improvements are understandably challenging to achieve. This makes the lack of a large effect unsurprising, as substantial performance gains between two assignments are likely unrealistic. Nonetheless, the observed positive, though not statistically significant, effect is reassuring.

Table 11: Second-stage results where the dependent variable is a continuous outcome on value-added scores for the second assignment.

Anderson-Rubin Test	**2SLS / Fuller(1) t-test**					
under F-distribution	**Method**	**Estimate**	**Std. Error**	**t-value**	**Pr($>|t|$)**	**Significance**
$F = 0.116$, p-value = 0.733	2SLS	0.027	0.080	0.342	0.733	
95% CI: [-0.136, 0.183]	Fuller(1)	0.028	0.079	0.357	0.721	
Notes: Significance codes: *** (p<0.001), ** (p<0.01), * (p<0.05), . (p < 0.1)						

5. Discussion

The findings from this study highlight the nuanced effects of teachers' viewing and use of the assignment reports in an intelligent tutoring system (ITS) on student assignment outcomes. It is important to note that the treatment effects which are estimated apply only for teachers who may be successfully encouraged to use the assignment report, and should be interpreted in the context of other coexisting teaching practices. While the overall impact of viewing the first assignment report on student outcomes on the next assignment was mixed, some key patterns emerged that enhance our understanding of ITS integration in teaching practices.

Impact on student starting and completion rates. The most robust finding in the study was that viewing of the first assignment report can positively impact the proportion of students that complete the next assignment. This effect, as validated by the Anderson-Rubin test, demonstrates its use as a monitoring tool to increase student participation in their work. Nonetheless, the unexpectedly high estimate from the Fuller(1) estimator suggests caution in interpreting the magnitude of this effect, as it may be inflated due to the weak instrument. While the impact on the proportion of students starting the next assignment was not statistically significant, this effect may be masked by coexisting teaching practices, such as initiating the assignment during class time.

Encouragement heterogeneity. An unplanned but noteworthy observation emerged regarding the heterogeneity in impact of encouragement on teachers' use of the assignment report. Teachers with a larger number of students were more likely to use the assignment report regardless of their assignment to the encouragement or control group. This suggests that these teachers may already rely on ITS features, such as reports, to manage the complexity of teaching larger classes effectively. Future studies using randomized encouragement designs to examine ITS features should consider excluding such teachers as a population of interest, as their baseline use of these tools may confound the estimation of treatment effects.

5.1. Implications for Teachers

The results support the conclusion that teachers who engage with the learning analytics reports provided by intelligent tutoring systems (ITS) achieve improved student outcomes, particularly in assignment completion rates. The observed increase in completion rates, combined with prior research on ITS improving next-problem correctness, suggests potential for cumulative benefits over time. Although the analysis revealed negligible impact on value-added scores for the second assignment—unsurprising given the challenge of achieving significant performance gains between two assignments—the positive, albeit statistically insignificant, effect points to the potential for sustained teacher use of assignment reports to drive incremental improvements in long-term student learning outcomes.

5.2. Contributions to Research

This study provides causal evidence of the benefits of teachers' engagement of the learning analytics report in ITS. By leveraging a large-scale randomized encouragement design, we were able to isolate the causal impact of report viewing on student outcomes, particularly on improving assignment completion rates. These findings contribute to the growing body of evidence demonstrating that teachers' use of ITS data can positively impact student outcomes. More importantly, this study underscores the critical role of teachers within the ITS ecosystem to enhance student learning in schools.

5.3. Limitations

The findings of this study are specific to ASSISTments and may not generalize well to other ITS platforms with different user interfaces or assignment report features. Additionally, the profiles of teachers and students using ASSISTments may differ from those using other ITS platforms, limiting the applicability of the results. Moreover, this study focuses solely on the impact of report viewing for the first assignment on student outcomes in the next assignment. Sustained use of the report is unlikely to produce similar gains over time but may help maintain outcomes.

Furthermore, due to the limitations of the randomized encouragement design, the findings are most relevant for teachers who can be successfully encouraged to view the assignment report (compliers). It is important to note that while we are unable to estimate the effect of report viewing for teachers who already use the report (always-takers) or those who choose not to use it (never-takers), this does not imply there is no effect. Rather, the study design does not allow us to detect these effects.

In addition, the instrument used in the encouragement condition was relatively weak, as indicated by low first-stage F-statistics. While the resulting estimates on completion rate remain valid under the Anderson–Rubin test, weak instrument bias may still affect both the magnitude and precision of the estimated local average treatment effect (LATE). Consequently, the estimated treatment effects should be interpreted with caution and understood as conservative given the weak instrumentation.

Finally, the study revealed heterogeneity in teachers' responsiveness to the encouragement — especially among those with more students — suggesting that these teachers tended to engage with assignment reports at higher rates regardless of encouragement assignment. In other words, they are likely overrepresented among always-takers. Their inclusion in the estimation sample may attenuate the estimated LATE, as it averages over subgroups with heterogeneous gains from treatment. The subsequent post-hoc analysis excluding these teachers yielded higher treatment effects, suggesting that the original estimates may have been biased downward due to selection on gains. As discussed in Angrist (2004), when treatment effects are heterogeneous and compliance is endogenous, the LATE may not generalize beyond the complier subgroup, and selection on gains can distort causal inference. This underscores the importance of carefully defining the estimation sample and targeting the encouragement to subpopulations with greater margin for behavioral change in future randomized encouragement designs.

5.4. Future Research

Building on the findings of this study, several areas for future research can be explored. The choice of encouragement in this study resulted in a valid but weak instrument. Future studies could test stronger modes of encouragement, such as email reminders with direct links to assignment reports, while excluding teachers with a large number of students from the study population. These adjustments may help produce more robust estimates of treatment effects in a randomized encouragement design.

Although this mode of encouragement significantly increased the proportion of overall teachers using the assignment report, the instrument's weakness suggests the need for alternative approaches. Targeting teachers experienced in ASSISTments with stronger encouragement strategies may provide additional insights. Future research could examine how impacts on student outcomes vary based on these teachers' usage profiles, considering factors such as whether assignments were primarily used as homework or classwork, or whether teachers focused on student-level details (e.g., individual performance) versus problem-level details (e.g., problem charts) when interacting with assignment data.

Finally, future studies could investigate other modes of teacher engagement with ITS, such as providing feedback on students' open responses, to further understand how different aspects of teacher involvement contribute to improved student outcomes. This would offer a more holistic perspective on the critical role of teachers within the ITS ecosystem.

6. Conclusion

This study provides causal evidence of the benefits of teachers' engagement with learning analytics reports in intelligent tutoring systems (ITS). Using a large-scale randomized encouragement design, we demonstrated that teachers who were successfully encouraged to view assignment reports positively influenced student assignment completion rates. Although the impact on value-added scores was negligible, the findings highlight the potential for incremental improvements through sustained use of the report.

These results underscore the critical role of teachers within the ITS ecosystem and emphasize the importance of designing ITS platforms that recognize the importance of the teacher's role in the ITS ecosystem in supporting student learning. Although the findings are specific to ASSISTments, they contribute to the growing body of research that advocates teacher engagement with ITS data as a key factor in improving student outcomes.

ACKNOWLEDGEMENTS

We would like to thank our many past and current funders including NSF (2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 2215842, 2341948, & 2153481); IES (R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, R305A120125, R305R220012, & R305T240029); GAANN (P200A120238, P200A180088, & P200A150306); DOE (U411B190024 S411 B210024, & S411B220024); ONR (N00014-18-1-2768); NIH (via a SBIR R44GM146483), Schmidt Futures, BMGF, CZI, Arnold, Hewlett, the Jaffe Foundation, and anonymous donors. None of the opinions expressed here are those of the funders.

REFERENCES

ALEVEN, V., BLANKESTIJN, J., LAWRENCE, L., NAGASHIMA, T., AND TAATGEN, N. 2022. A dashboard to support teachers during students' self-paced AI-supported problem-solving practice. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, I. Hilliger, P. J. Muñoz-Merino, T. De Laet, A. Ortega-Arranz, and T. Farrell, Eds. Vol. 13450. Springer International Publishing, Cham, 16–30. Series Title: Lecture Notes in Computer Science.

ANDREWS, I. AND ARMSTRONG, T. B. 2017. Unbiased instrumental variables estimation under known first-stage sign: Unbiased IV estimation. *Quantitative Economics 8,* 2 (July), 479–503.

ANDREWS, I., STOCK, J. H., AND SUN, L. 2019. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics 11,* 1 (Aug.), 727–753.

ANGRIST, J. AND KOLESÁR, M. 2024. One instrument to rule them all: The bias and coverage of just-ID IV. *Journal of Econometrics 240,* 2 (Mar.), 105398.

ANGRIST, J. AND LAVY, V. 2009. The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review 99,* 4 (Aug.), 1384–1414.

ANGRIST, J. D. 2001. Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business & Economic Statistics 19,* 1 (Jan.), 2–28.

ANGRIST, J. D. 2004. Treatment effect heterogeneity in theory and practice. *The Economic Journal 114,* 494 (Mar.), C52–C83.

ANGRIST, J. D., IMBENS, G. W., AND RUBIN, D. B. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91,* 434 (June), 444–455.

ANGRIST, J. D. AND PISCHKE, J.-S. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ.

ASSISTMENTS. 2024. Explore the assignment report. The ASSISTments Foundation. https://new.assistments.org/resources/explore-the-assignment-report. Accessed March 18, 2025.

BAKER, H. D. 2003. Teaching with ALEKS. ALEKS Corporation. https://www.aleks.com/manual/pdf/teaching.pdf. Accessed March 18, 2025.

BAKER, R. S. 2016. Stupid Tutoring Systems, Intelligent Humans. *International Journal of Artificial Intelligence in Education 26,* 2 (June), 600–614.

BOUND, J., JAEGER, D. A., AND BAKER, R. M. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association 90,* 430 (June), 443.

BRADLOW, E. 1998. Encouragement designs: An approach to self-selected samples in an experimental design. *Marketing Letters 9,* 4, 383–391.

BURGESS, S., SMALL, D. S., AND THOMPSON, S. G. 2017. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research 26,* 5 (Oct.), 2333–2355.

CARNEGIE LEARNING. 2025. Getting started with MATHia reports. https://support.carnegielearning.com/help-center/math/mathia-reports/general/article/getting-started-mathia-reports. Accessed 18 Mar 2025.

FENG, M., HUANG, C., AND COLLINS, K. 2023. Technology-based support shows promising long-term impact on math learning: Initial results from a randomized controlled trial in middle schools.

GINSBURG, A. AND SMITH, M. S. 2016. Do randomized controlled trials meet the "gold standard"? A study of the usefulness of RCTs in the What Works Clearinghouse. Tech. rep., American Enterprise Institute. https://www.carnegiefoundation.org/wp-content/uploads/2016/03/Do-randomized-controlled-trials-meet-the-gold-standard.pdf. Accessed 18 Mar 2025.

GREENE, W. 2018. *Econometric analysis*, Eighth ed. Pearson, New York, NY.

HAHN, J., HAUSMAN, J., AND KUERSTEINER, G. 2004. Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. *The Econometrics Journal 7,* 1 (June), 272–306.

HAYASHI, F. 2000. *Econometrics*. Princeton University Press, Princeton.

HEFFERNAN, N. T. AND HEFFERNAN, C. L. 2014. 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 (Dec.), 470–497.

HELSABECK, N. P., JUSTICE, L. M., AND LOGAN, J. A. R. 2022. Assessing fidelity of implementation to a technology-mediated early intervention using process data. *Journal of Computer Assisted Learning 38,* 2 (Apr.), 409–421.

HOLSTEIN, K., HONG, G., TEGENE, M., MCLAREN, B. M., AND ALEVEN, V. 2018. 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*. ACM, Sydney New South Wales Australia, 79–88.

HOLSTEIN, K., MCLAREN, B. M., AND ALEVEN, V. 2017. Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, Vancouver British Columbia Canada, 257–266.

KANG, H., JIANG, Y., ZHAO, Q., AND SMALL, D. S. 2021. ivmodel: An R package for inference and sensitivity analysis of Instrumental variables models with one endogenous variable. *Observational Studies 7,* 2, 1–24.

KEANE, M. P. AND NEAL, T. 2024. A practical guide to weak instruments. *Annual Review of Economics 16,* 1 (Aug.), 185–212.

KELLER, T. AND SZAKÁL, P. 2021. Not just words! Effects of a light-touch randomized encouragement intervention on students' exam grades, self-efficacy, motivation, and test anxiety. *PLOS ONE 16,* 9 (Sept.), e0256960.

KELLY, K., HEFFERNAN, N., HEFFERNAN, C., GOLDMAN, S., PELLEGRINO, J., AND SOFFER GOLDSTEIN, D. 2013. Estimating the effect of web-based homework. In *Artificial Intelligence in Education*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Vol. 7926. Springer Berlin Heidelberg, Berlin, Heidelberg, 824–827. Series Title: Lecture Notes in Computer Science.

KOCHMAR, E., VU, D. D., BELFER, R., GUPTA, V., SERBAN, I. V., AND PINEAU, J. 2020. Automated personalized feedback improves learning gains in an intelligent tutoring system. In *Artificial Intelligence in Education*, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Vol. 12164. Springer International Publishing, Cham, 140–146. Series Title: Lecture Notes in Computer Science.

KOEDINGER, K. R. AND ALEVEN, V. 2016. An interview reflection on "Intelligent tutoring goes to school in the big city". *International Journal of Artificial Intelligence in Education 26,* 1 (Mar.), 13–24.

KOEDINGER, K. R., ANDERSON, J. R., HADLEY, W. H., AND MARK, M. A. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education 8*, 30–43.

LEE, M., SIEDAHMED, A., AND HEFFERNAN, N. 2024. Expert features for a student support recommendation contextual bandit algorithm. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, Kyoto Japan, 864–870.

LEY, T., TAMMETS, K., PISHTARI, G., CHEJARA, P., KASEPALU, R., KHALIL, M., SAAR, M., TUVI, I., VÄLJATAGA, T., AND WASSON, B. 2023. Towards a partnership of teachers and intelligent learning technology: A systematic literature review of model-based learning analytics. *Journal of Computer Assisted Learning 39,* 5 (Oct.), 1397–1417.

LOUSDAL, M. L. 2018. An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology 15,* 1 (Jan.), 1.

MA, W., ADESOPE, O. O., NESBIT, J. C., AND LIU, Q. 2014. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology 106,* 4 (Nov.), 901–918.

MARTENS, E. P., PESTMAN, W. R., DE BOER, A., BELITSER, S. V., AND KLUNGEL, O. H. 2006. Instrumental variables: Application and limitations. *Epidemiology 17,* 3 (May), 260–267.

MARTIN, F., SUN, T., AND WESTINE, C. D. 2020. A systematic review of research on online teaching and learning from 2009 to 2018. *Computers & Education 159*, 104009.

MCGRAW HILL EDUCATION. 2025. Aleks reporting. https://www.mheducation.com/prek-12/program/microsites/MKTSP-GAB02M0.html#reporting. Accessed March 18, 2025.

NYE, B. D. 2014. Barriers to ITS adoption: A systematic mapping study. In *Intelligent Tutoring Systems*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, A. Kobsa, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, D. Terzopoulos, D. Tygar, G. Weikum, S. Trausan-Matu, K. E. Boyer, M. Crosby, and K. Panourgia, Eds. Vol. 8474. Springer International Publishing, Cham, 583–590. Series Title: Lecture Notes in Computer Science.

OECD. 2023. *PISA 2022 Results (Volume II): Learning During – and From – Disruption*. PISA. OECD, Paris.

PALOYO, A. R., ROGAN, S., AND SIMINSKI, P. 2016. The effect of supplemental instruction on academic performance: An encouragement design experiment. *Economics of Education Review 55*, 57–69.

PANE, J. F., GRIFFIN, B. A., MCCAFFREY, D. F., AND KARAM, R. 2014. Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis 36,* 2 (June), 127–144.

PANTELIMON, F.-V., BOLOGA, R., TOMA, A., AND POSEDARU, B.-S. 2021. The evolution of AI-driven educational systems during the covid-19 pandemic. *Sustainability 13,* 23 (Dec.), 13501.

PATIKORN, T. AND HEFFERNAN, N. T. 2020. Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*. ACM, Virtual Event USA, 115–124.

RAZZAQ, L. AND HEFFERNAN, N. 2009. To tutor or not to tutor: That is the question. In *Proceedings of the 2009 Artificial Intelligence in Education Conference*. IOS Press, Brighton, UK, 457–464.

SCHOCHET, P. Z. AND CHIANG, H. S. 2011. Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioral Statistics 36,* 3 (June), 307–345.

SHI, X. 2024. Lecture 11: Weak instruments. Econ 715 Lecture Notes, University of Wisconsin-Madison. https://users.ssc.wisc.edu/ xshi/econ715/Lecture_11_WeakIV.pdf. Accessed 18 Mar 2025.

STAIGER, D. AND STOCK, J. H. 1997. Instrumental variables regression with weak instruments. *Econometrica 65,* 3 (May), 557–586.

STOCK, J. H. AND YOGO, M. 2005. Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, D. W. K. Andrews and J. H. Stock, Eds. Cambridge University Press, Cambridge, 80–108.

USERFLOW. 2023. A/B testing. Userflow Documentation. https://docs.userflow.com/docs/guides/ab-testing. Accessed 30 Nov 2024.

USERFLOW. 2024. Userflow Documentation. Userflow Documentation. https://docs.userflow.com/docs. Accessed 30 Nov 2024.

VANACORE, K., DIETER, K., HURWITZ, L., AND STUDWELL, J. 2021. Longitudinal clusters of online educator portal access: Connecting educator behavior to student outcomes. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. ACM, Irvine CA USA, 540–545.

VANACORE, K., GURUNG, A., SALES, A., AND HEFFERNAN, N. T. 2024. The effect of assistance on gamers: Assessing the impact of on-demand hints & feedback availability on learning for students who game the system. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, Kyoto Japan, 462–472.

WEST, S. G., DUAN, N., PEQUEGNAT, W., GAIST, P., DES JARLAIS, D. C., HOLTGRAVE, D., SZAPOCZNIK, J., FISHBEIN, M., RAPKIN, B., CLATTS, M., AND MULLEN, P. D. 2008. Alternatives to the randomized controlled trial. *American Journal of Public Health 98,* 8 (Aug.), 1359–1366.

WHAT WORKS CLEARINGHOUSE. 2022. What Works Clearinghouse procedures and standards handbook, version 5.0. Condition of Education. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). This report is available on the What Works Clearinghouse website at https://ies.ed.gov/ncee/wwc/Handbooks.

XHAKAJ, F., ALEVEN, V., AND MCLAREN, B. M. 2016. How teachers use data to help students learn: Contextual inquiry for the design of a dashboard. In *Adaptive and Adaptable Learning*, K. Verbert, M. Sharples, and T. Klobučar, Eds. Vol. 9891. Springer International Publishing, Cham, 340–354. Series Title: Lecture Notes in Computer Science.

XHAKAJ, F., ALEVEN, V., AND MCLAREN, B. M. 2017. Effects of a teacher dashboard for an intelligent tutoring system on teacher knowledge, lesson planning, lessons and student learning. In *Data Driven Approaches in Digital Education*, E. Lavoué, H. Drachsler, K. Verbert, J. Broisin, and M. Pérez-Sanagustín, Eds. Vol. 10474. Springer International Publishing, Cham, 315–329. Series Title: Lecture Notes in Computer Science.

XIAO, Z., HAUSER, O., KIRKWOOD, C., LI, D. Z., FORD, T., AND HIGGINS, S. 2024. Uncovering individualised treatment effects for educational trials. *Scientific Reports 14,* 1 (Sept.), 22606.

ZHOU, G., YANG, X., AZIZSOLTANI, H., BARNES, T., AND CHI, M. 2020. Improving student-system interaction through data-driven explanations of hierarchical reinforcement learning induced pedagogical policies. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Genoa Italy, 284–292.