Pre-Analysis Plan

NUMI, Mastery Learning, and Short-Duration Math Practice

April 9, 2026

Abstract

This pre-analysis plan describes the analysis of an experiment conducted in middle school math classes in Hamilton County Department of Education. Students were randomized to (i) receive access to the AI tutor NUMI or not, (ii) complete the activity under mastery-learning rules or not, and (iii) practice one of two topic bundles. One week later, students completed a delayed test containing one question corresponding to each exercise type for each topic bundle. The central feature of the design is that each student has both a practiced-topic outcome and an unpracticed-topic outcome, allowing the main estimand to compare each student with themselves.

The primary outcome is the within-student difference between delayed-test performance on the practiced topic and delayed-test performance on the unpracticed topic. The main question is whether this practice effect differs across the four treatment groups defined by mastery and AI assignment. The design is motivated by the hypothesis that AI tutoring will be more effective under mastery learning, because mastery creates stronger incentives to resolve mistakes rather than skip ahead. At the same time, mastery may slow students down, reducing the probability that they reach later exercises and the exit ticket. Accordingly, the analysis distinguishes between total effects on learning, effects on progression through the practice sequence, and effects on AI take-up and post-error engagement.

The plan also pre-specifies decompositions of delayed-test effects by exercise position, balance and attrition checks, analyses of delayed-test missingness, progression and mastery-gate outcomes, AI-usage outcomes, prompt-level interaction analyses, exploratory robustness checks related to platform slowness and low-effort delayed-test completion, and heterogeneity by grade, prior ability, and selected exploratory post-treatment measures.

1 Study Overview

This study evaluates whether AI tutoring and mastery learning improve student learning during a short in-class math practice session.

The intervention was conducted during a 50-minute math class. Students logged into an online platform and were assigned to practice one of two topic bundles. Each topic bundle contained two sets of content, where each set consisted of one video and one exercise. Students then completed an exit ticket if they reached that point in the activity.

The experiment has three randomized dimensions:

1. **AI assignment:** access to NUMI versus no NUMI;
2. **Mastery assignment:** mastery-learning version versus non-mastery version;
3. **Topic assignment:** Topic A versus Topic B, with the topic content varying by grade.

Approximately one week later, students took a delayed test. The test included one question corresponding to each exercise type from each topic bundle. This makes it possible to compare, within student, performance on practiced material versus unpracticed material.

2 Intervention Details

2.1 Mastery versus non-mastery

In the **mastery** condition, students had to:

- watch at least one minute of the current video, and
- answer three exercise questions in a row correctly

before moving on to the next set, and again before accessing the exit ticket.

In the **non-mastery** condition, students could move more freely through the content. They had to attempt at least one question, but they could skip around among videos and exercises and proceed to the exit ticket without satisfying the mastery threshold.

2.2 AI tutor versus no AI

Students assigned to the AI condition had access to **NUMI**, an AI math tutor. NUMI offered several kinds of support:

a. **Get-started hints:** students could request structured hints about the first steps of the solution;

b. **Forced post-error walkthroughs:** after some mistakes, students were required to walk through the solution with NUMI;

c. **Step explanations:** students could click on specific steps of the solution to request more detail or intuition;

d. **Open-ended interaction:** students could type math-related questions or short responses.

NUMI's structured interactions included three main prompt types:

1. a yes/no understanding check,
2. a button-based A/B choice prompt,
3. a short open-ended prompt.

Students not assigned to AI could see solutions after mistakes but did not receive tutoring support.

2.3 Motivation for the design

A central motivation for introducing mastery learning is that, without mastery requirements, students may have little incentive to engage with the AI tutor after making mistakes because they can simply move on or skip ahead. Under mastery learning, mistakes become more consequential because students must answer three questions in a row correctly in order to progress. This may increase the take-up and effectiveness of AI assistance.

At the same time, mastery may slow students down. Every incorrect answer resets the sequence of correct answers required for progression, which may reduce the probability that students reach the second exercise or the exit ticket. The analysis is therefore organized around a tradeoff between *depth of engagement* and *breadth of completed material*.

3 Research Questions and Hypotheses

3.1 Primary research question

Does the effect of practice on delayed-test performance differ across the four treatment groups defined by mastery and AI assignment?

The four treatment groups are:

1. Mastery + AI
2. Non-mastery + AI
3. Mastery + no AI
4. Non-mastery + no AI

3.2 Primary hypothesis

The combination of mastery learning and AI tutoring will generate the largest gain in delayed-test performance on practiced relative to unpracticed material.

3.3 Mechanism hypothesis

AI tutoring will be more effective under mastery learning because mastery creates stronger incentives to resolve mistakes. In the absence of mastery, students can move on or skip ahead without engaging deeply with the AI tutor.

3.4 Tradeoff hypothesis

Mastery learning may improve depth of learning on earlier material while reducing the probability that students reach later material. As a result, treatment effects may differ for delayed-test questions linked to the first versus second exercise.

3.5 AI engagement hypothesis

A key implementation challenge is whether students actually engage with NUMI. NUMI was designed to lower friction and encourage engagement through button-based responses, forced interactions after some mistakes, optional hints, and optional explanation features. However, open-ended prompts may slow students down or discourage response. The analysis therefore pre-specifies AI usage and prompt-level process outcomes.

3.6 Platform performance hypothesis

Platform slowness may have reduced the quality of the intervention itself, especially for students assigned to AI, whose treatment experience depended more directly on interactive responsiveness. We therefore expect estimated AI effects to be weaker for students exposed to substantial website slowness during the practice session.

3.7 Outcome-quality hypothesis

Some students may not have taken the delayed test seriously. Very low-effort test completion may reduce the quality of the outcome measure and attenuate estimated treatment effects. We therefore plan exploratory robustness analyses excluding students who appear not to have seriously engaged with the delayed test.

4 Estimands

4.1 Primary estimand

For each treatment arm, the primary estimand is the average within-student difference between delayed-test performance on the practiced topic and delayed-test performance on the unpracticed topic.

4.2 Comparative estimands

The key treatment comparisons are:

- AI effect among non-mastery students;
- AI effect among mastery students;
- mastery effect among no-AI students;
- mastery effect among AI students;
- complementarity between AI and mastery;
- difference between Mastery + AI and Non-mastery + no AI.

4.3 Secondary estimands

Secondary estimands include treatment effects on:

- progression through the practice sequence,
- mastery-gate completion,
- delayed-test missingness,
- AI take-up and usage,
- post-error recovery,
- exercise-specific delayed-test outcomes.

4.4 Exploratory robustness estimands

Exploratory robustness analyses will examine:

- whether treatment effects differ for students exposed to slow website periods,
- whether excluding students exposed to severe slowness changes estimated treatment effects,
- whether excluding students who appear not to have seriously engaged with the delayed test changes estimated treatment effects.

5 Outcome Definitions

5.1 Primary delayed-test outcome

Let:

- Y_i^{P1} = score on the delayed-test question corresponding to practiced exercise 1,
- Y_i^{P2} = score on the delayed-test question corresponding to practiced exercise 2,
- Y_i^{NP1} = score on the delayed-test question corresponding to unpracticed exercise 1,
- Y_i^{NP2} = score on the delayed-test question corresponding to unpracticed exercise 2.

The primary outcome is:

$$D_i = (Y_i^{P1} + Y_i^{P2}) - (Y_i^{NP1} + Y_i^{NP2}). \tag{1}$$

If each delayed-test item is coded as correct/incorrect, then D_i ranges from -2 to 2.

5.2 Exercise-specific delayed-test outcomes

Because mastery may reduce the probability that students reach the second exercise, the analysis will separately examine delayed-test effects by exercise position:

$$D_{i1} = Y_i^{P1} - Y_i^{NP1}, \tag{2}$$

$$D_{i2} = Y_i^{P2} - Y_i^{NP2}. \tag{3}$$

These are key pre-specified secondary outcomes.

5.3 Additional delayed-test outcomes

Secondary delayed-test outcomes include:

- practiced-topic score: $Y_i^{P1} + Y_i^{P2}$,
- unpracticed-topic score: $Y_i^{NP1} + Y_i^{NP2}$,
- total delayed-test score,
- question-level delayed-test correctness.

5.4 Progression outcomes

Across all students, progression outcomes include:

- watched first video for at least one minute,
- attempted first exercise,
- number of questions attempted from first exercise,
- number of questions correct from first exercise,
- completed first set,
- reached second video,
- attempted second exercise,
- number of questions attempted from second exercise,
- number of questions correct from second exercise,
- completed second set,
- reached exit ticket,
- exit ticket score,
- submitted exit ticket,
- total questions attempted,
- total time on platform,
- total content completed.

5.5 Mastery-specific process outcomes

Among students assigned to mastery, the following outcomes will be analyzed:

- cleared first mastery gate,
- cleared second mastery gate,
- cleared both gates,
- attempts to clear first gate,
- attempts to clear second gate,
- time to clear first gate,
- time to clear second gate.

Precise coding definitions for these milestones will be documented in a data appendix.

5.6 Post-error recovery outcomes

Exploratory post-error outcomes include:

- whether the student eventually answers correctly after the first mistake,
- whether the student clears the gate after the first mistake,
- number of attempts from first mistake to next correct answer,
- time from first mistake to next correct answer,
- correctness on the next attempted question after the first mistake.

5.7 AI engagement outcomes

Among AI-assigned students, AI-engagement outcomes include:

- any use of NUMI,
- time from login to first NUMI use
- any NUMI use before the first mistake
- any NUMI use after the first mistake
- NUMI interactions per attempted question
- total number of NUMI interactions,
- total time interacting with NUMI,
- any use of "help get me started,"
- number of hint uses,
- any typed open-ended message,
- number of typed messages,
- number of forced walkthroughs triggered,
- number of forced walkthroughs completed,
- number of optional step-explanation clicks,
- share of interactions using buttons rather than typed responses.

AI-assisted evaluation of open-ended NUMI chat data. In addition to headline measures of AI take-up, we will analyze the content of open-ended NUMI chat exchanges using a frozen LLM-based evaluation agent. A chat episode will be defined as the chat initiated by a student typed message and followed by the corresponding NUMI response sequence. The goal is to characterize what students ask NUMI for and what kinds of support NUMI provides. At a high level, the agent will classify episodes into a small set of pre-specified categories, such as conceptual explanation, procedural guidance, answer checking, off-task interaction, or other implementation-relevant content. These transcript-based measures are intended as mechanism and implementation outcomes rather than primary learning outcomes.

Before full-sample analysis, we will finalize a compact coding scheme and then hold fixed the evaluation prompt, model configuration, and transcript-preprocessing rules. The evaluation agent will be blinded to delayed-test outcomes and other downstream outcome data. Validated transcript measures will be aggregated to the student level and used in descriptive analyses among AI-assigned students. For example, we may examine whether mastery changes the share of chats that are conceptual rather than procedural, the share that appear off-task, or the share that end with a clear actionable next step. We may also report descriptive associations between these aggregated chat measures and progression or delayed-test outcomes. Because chat content is jointly shaped by student behavior and NUMI responses, these analyses will be interpreted as descriptive evidence on implementation and mechanisms rather than as causal estimates.

5.8 Prompt-level outcomes

At the prompt/event level, outcomes include:

- response indicator,
- response latency,
- nonresponse,
- skip behavior,
- continuation to the next prompt,
- continuation to the next problem.

Prompt types will be coded as:

1. yes/no understanding prompt,
2. A/B button-choice prompt,
3. short open-ended prompt.

5.9 Missingness outcomes

Missingness outcomes include:

- missing delayed test in week 2,
- missing week-1 intervention session,
- observed in week 2 conditional on assignment,
- attendance in week 1 and week 2.

5.10 Platform slowness measures

Exploratory measures of platform performance will be constructed from timestamped activity logs. Candidate measures include:

- average wait time between student request and next platform response,
- median wait time between requests,
- fraction of waits exceeding 10 seconds,
- an indicator for exposure to a "slow period" defined by generally having to wait more than 10 seconds between requests,
- day-of-experiment indicators, especially day 1–2 versus day 3–4.

Because the most reliable slowness measure may depend on the available logs, multiple operationalizations may be examined in the appendix. A leading candidate definition is an indicator for whether a student was active during a period with average waits above 10 seconds.

5.11 Low-effort delayed-test outcomes

Exploratory low-effort delayed-test measures include:

- answered no delayed-test questions,
- completed the delayed test in less than 3 minutes,
- indicator for satisfying either of the above.

6 Sample Definitions

6.1 Full randomized sample

All randomized students in participating middle school classes.

6.2 Week-1 intervention sample

Students present in week 1 and observed to log into the platform.

6.3 Week-2 delayed-test sample

Students observed on the delayed test in week 2, whether or not they attended week 1.

6.4 Main analysis sample

Students with:

- treatment assignment,
- topic assignment,
- delayed-test data for both practiced-topic and unpracticed-topic questions.

6.5 Students absent in week 1 but present in week 2

Some students missed the practice session but took the delayed test. These students will be excluded from the main causal analysis because they did not receive the intended practice treatment. They will, however, be described in sample-count tables and may be used in supplemental descriptive analyses as an added non-exposed comparison group.

6.6 Exploratory restricted samples

Two exploratory restricted samples will be examined as robustness checks.

First, because website performance was slower during some periods of data collection, we will construct measures of platform slowness based on the time students experienced the site. The experiment took place over four days, and performance issues were mitigated after the second day. We will therefore examine whether treatment effects differ for students who used the site during slower periods, and whether excluding students exposed to substantial slowness changes the estimated effects.

Second, we will construct an exploratory restricted delayed-test sample that excludes students who appear not to have taken the delayed test seriously. Candidate definitions include students who answered no delayed-test questions or completed the delayed test in less than three minutes. These exclusions will not define the default main sample unless they materially affect interpretation.

7 Balance, Compliance, and Attrition Checks

7.1 Sample counts

The first table will report sample counts overall and by treatment arm:

- total randomized,
- week-1 attendance,
- week-1 platform login,
- week-2 delayed-test observation,
- main analysis sample,
- week-2-only students.

Counts will also be shown by grade where feasible.

7.2 Balance tests

Pre-treatment covariates will be tested for balance across treatment arms. Subject to availability, these covariates include:

- grade,
- gender,
- prior achievement / baseline ability,
- school or classroom indicators,
- any clearly pre-treatment administrative variables.

The baseline balance regression is:

$$X_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i, \tag{4}$$

where X_i is a pre-treatment covariate and δ_g are grade fixed effects.

7.3 Topic difficulty balance

A key identifying assumption of the within-student design is that, conditional on grade, the two topic bundles are equally difficult. To test this, we regress the unpracticed-topic delayed-test score on topic assignment with grade fixed effects:

$$Y_{ig}^{NP} = \alpha + \beta \, \mathrm{TopicA}_i + \delta_g + \varepsilon_{ig} \tag{5}$$

where $Y_{ig}^{NP} = Y_i^{NP1} + Y_i^{NP2}$ is the sum of scores on unpracticed questions. This is estimated first on the full randomized sample as a test of whether assignment was fair, and then on the main analysis sample as a sensitivity check for whether differential attrition has broken that balance for the students who identify the primary estimates. No additional controls are included in either specification; the goal is to test the null $E[\text{TopicA}_i \mid \varepsilon_{ig}] = 0$ cleanly. A precisely estimated $\hat{\beta} \neq 0$ in the analysis sample would indicate that topic difficulty is confounding D_i, and the primary specification will be augmented with a TopicA_i indicator as a robustness check.

7.4 Attendance and participation

Treatment differences in week-1 attendance or login will be checked using:

$$AttendWeek1_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i. \tag{6}$$

7.5 Delayed-test attrition

A key check is whether missing delayed-test outcomes are correlated with treatment:

$$MissingWeek2_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i. \tag{7}$$

A secondary specification will add practice-topic assignment:

$$MissingWeek2_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \beta_4 TopicA_i + \delta_g + \varepsilon_i. \tag{8}$$

If delayed-test missingness is meaningfully correlated with treatment, this will be reported prominently and sensitivity analyses will be added.

7.6 Low-effort delayed-test balance

We will also test whether low-effort delayed-test behavior differs by treatment assignment:

$$LowEffortTest_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i. \tag{9}$$

8 Primary Estimation Strategy

8.1 Primary student-level difference model

The preferred primary specification is:

$$D_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i, \tag{10}$$

where:

- D_i is the practiced-minus-unpracticed delayed-test score,
- M_i is an indicator for mastery assignment,
- A_i is an indicator for AI assignment,
- δ_g are grade fixed effects.

Interpretation:

- α: practice effect in the non-mastery, no-AI group,
- β_1: additional effect of mastery,
- β_2: additional effect of AI,
- β_3: complementarity between mastery and AI.

8.2 Key hypothesis tests

The following linear hypotheses will be pre-specified:

$$\begin{aligned}
\text{AI effect among non-mastery students:} &\quad H_0 : \beta_2 = 0, \\
\text{Mastery effect among no-AI students:} &\quad H_0 : \beta_1 = 0, \\
\text{AI effect among mastery students:} &\quad H_0 : \beta_2 + \beta_3 = 0, \\
\text{Mastery effect among AI students:} &\quad H_0 : \beta_1 + \beta_3 = 0, \\
\text{Mastery + AI vs. Non-mastery + no AI:} &\quad H_0 : \beta_1 + \beta_2 + \beta_3 = 0, \\
\text{Complementarity:} &\quad H_0 : \beta_3 = 0.
\end{aligned}$$

8.3 Equivalent student fixed-effects specification

As an equivalent formulation, the main estimand can also be represented in student-topic format:

$$Y_{it} = \alpha_i + \gamma P_{it} + \beta_1(P_{it} \times M_i) + \beta_2(P_{it} \times A_i) + \beta_3(P_{it} \times M_i \times A_i) + \lambda_{gt} + \varepsilon_{it}, \tag{11}$$

where:

- Y_{it} is student i's score on topic t,
- $P_{it} = 1$ if topic t is the practiced topic,
- α_i is a student fixed effect,
- λ_{gt} are grade-by-topic fixed effects.

This specification uses each student as their own control. The difference-score model remains the primary presentation because it is simpler to interpret.

9 Exercise-Specific Learning Analyses

Because the delayed test includes one question corresponding to each exercise, and because mastery may reduce exposure to the second exercise, the learning analysis will explicitly distinguish between exercise 1 and exercise 2.

9.1 Exercise 1-specific practice effect

$$D_{i1} = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i. \tag{12}$$

9.2 Exercise 2-specific practice effect

$$D_{i2} = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i. \tag{13}$$

9.3 Interpretation

These two exercise-specific outcomes are central to interpretation, not merely robustness checks. If mastery deepens learning on the first exercise but reduces the probability of reaching the second, then treatment effects may be positive for D_{i1} and smaller or negative for D_{i2}. The primary combined outcome D_i should therefore be understood as the net effect across these potentially offsetting channels.

10 Secondary Delayed-Test Analyses

10.1 Question-level delayed-test model

Question-level correctness will also be analyzed:

$$Y_{ij} = \alpha_i + \beta_1 PracticeQuestion_{ij} + \beta_2 (PracticeQuestion_{ij} \times M_i) + \beta_3 (PracticeQuestion_{ij} \times A_i) + \beta_4 (PracticeQuestio \tag{14}$$

with standard errors clustered at the student level.

10.2 Total delayed-test score

As a secondary outcome:

$$TotalScore_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i. \tag{15}$$

This outcome mixes practiced and unpracticed content and is therefore secondary to the within-student difference estimand.

11 Progression and Process Analyses

11.1 Progression outcomes across all students

For outcomes that are well-defined for all students:

$$Y_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i. \tag{16}$$

Key progression outcomes include:

- reached first video
- reached first exercise
- attempted first exercise
- first exercise score (number of questions correct/number of questions attempted)
- completed first exercise
- reached second video,
- attempted second exercise,
- completed second set,
- second exercise score (number of questions correct/number of questions attempted)
- completed second exercise
- reached exit ticket,
- exit ticket score
- submitted exit ticket,
- total questions attempted,
- total time on platform.

These outcomes capture the breadth side of the depth-versus-breadth tradeoff.

11.2 Mastery-specific process outcomes

Among students assigned to mastery:

$$Y_i = \alpha + \beta A_i + \delta_g + \varepsilon_i. \tag{17}$$

The most important mastery-specific outcomes are:

- cleared first gate,
- cleared second gate,
- cleared both gates,
- attempts to first gate,
- time to first gate,
- attempts to second gate,
- time to second gate.

11.3 Post-error recovery

Conditional on making at least one mistake:

$$Y_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i. \tag{18}$$

These analyses are exploratory because they condition on a post-randomization event.

12 AI Engagement and Mechanism Analyses

12.1 AI take-up

Among AI-assigned students, the first question is whether students used NUMI at all and how intensively they used it.

To test whether mastery increases demand for AI:

$$Usage_i = \alpha + \beta M_i + \delta_g + \varepsilon_i, \tag{19}$$

estimated among students assigned to AI.

As an additional mechanism analysis among AI-assigned students, we will descriptively examine the timing of NUMI use relative to the student's first mistake, classifying whether first NUMI use occurs before the first mistake, after the first mistake, or not at all. This directly relates to the hypothesis that mastery increases the incentive to use AI to resolve mistakes rather than move on. These timing analyses will be interpreted descriptively rather than causally because they depend on post-randomization behavior.

12.2 Prompt-type friction

A central concern is that open-ended prompts may slow students down or discourage interaction. Prompt-level analyses will estimate:

$$Y_e = \alpha + \beta_1 OpenEnded_e + \beta_2 ABPrompt_e + \gamma_i + \delta_g + \varepsilon_e, \tag{20}$$

where the omitted category is a yes/no prompt, and Y_e may be response latency, nonresponse, skip, or continuation.

Because prompt type is not independently randomized, these analyses will be interpreted descriptively rather than causally.

12.3 Forced walkthrough analyses

Among AI-assigned students who make at least one mistake, the analysis will summarize and model:

- completed first forced walkthrough,
- time spent in first forced walkthrough,
- number of prompts answered,
- whether the student skipped after the walkthrough,
- next-question correctness,
- eventual first-gate completion after the first mistake.

12.4 Usage-learning associations

Among AI-assigned students, descriptive associations between usage intensity and outcomes such as gate completion, reaching the exit ticket, and delayed-test practice gain will also be reported. These will be explicitly labeled descriptive and not causal.

13 Exploratory Robustness Analyses: Platform Slowness and Test Effort

13.1 Platform slowness during the intervention

Website performance varied during the experimental period, especially during earlier classes when the site slowed as more students used it simultaneously. Because this may have affected the user experience—particularly for students assigned to AI, whose experience depended more directly on interactive responsiveness—we will conduct exploratory analyses of treatment-effect heterogeneity and sample restrictions based on measured slowness.

The primary approach will be to construct student-level measures of platform slowness using timestamped activity logs. Candidate measures include:

- average wait time between user action and next platform response,
- median wait time between requests,
- fraction of requests associated with wait times greater than 10 seconds,
- an indicator for exposure to a slow period defined by sustained average waits above a prespecified threshold,
- day-of-experiment indicators, especially whether the student participated on day 1–2 versus day 3–4.

Because the exact slowness measure may depend on what can be reliably recovered from the logs, we will examine multiple operationalizations in the appendix. A leading candidate definition is an indicator for whether the student was using the site during a period in which average waits between requests exceeded 10 seconds.

Because student-level realized wait times may partly reflect the student's own usage pattern as well as underlying platform performance, a preferred robustness measure, when the logs permit it, will be a leave-one-out or common-environment slowness measure based on other students active in the same class period or narrow time window. This provides a cleaner proxy for platform conditions that are plausibly external to the individual student.

We will use these measures in three ways.

First, we will describe the distribution of slowness by day, class period, treatment arm, and grade.

Second, we will estimate whether treatment effects differ for students exposed to slower site performance using interaction models of the form:

$$D_i = \alpha+\beta_1 M_i+\beta_2 A_i+\beta_3(M_i\times A_i)+\beta_4 Slow_i+\beta_5(M_i\times Slow_i)+\beta_6(A_i\times Slow_i)+\beta_7(M_i\times A_i\times Slow_i)+\delta_g+\varepsilon_i. \tag{21}$$

Third, we will estimate the main treatment effects on progressively restricted samples that exclude students exposed to high-slowness periods. Candidate restrictions include:

- excluding students whose average wait time exceeds a threshold,
- excluding students whose fraction of waits above 10 seconds exceeds a threshold,
- excluding students observed during identified slow periods,
- excluding students observed during experimental days 1–2.

These analyses are exploratory. The default main specification will use the full sample. If the results appear highly sensitive to slowness exposure, we will report this prominently and may present restricted-sample estimates in the main text alongside the full-sample estimates.

13.2 Low-effort delayed-test completion

We will also conduct exploratory robustness analyses excluding students whose delayed-test behavior suggests that they did not take the assessment seriously.

Candidate low-effort definitions include:

- answering no delayed-test questions,
- completing the delayed test in less than three minutes,
- both of the above.

The primary delayed-test analysis will initially retain all observed students. We will then estimate robustness specifications excluding low-effort test takers:

$$D_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i \tag{22}$$

on the restricted samples.

We will also report whether the incidence of low-effort delayed-test behavior differs by treatment assignment:

$$LowEffortTest_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i. \tag{23}$$

These exclusions are exploratory and will not define the default main sample unless they materially affect interpretation. If they do, we will present both the full-sample and restricted-sample estimates in the main text.

13.3 Interpretation

These two exploratory analyses address distinct concerns. Platform slowness may have reduced the quality of the intervention itself, especially in the AI condition, thereby attenuating treatment effects. Low-effort delayed-test completion may instead reduce the quality of the outcome measure. For transparency, both issues will be examined in appendix robustness analyses, and any material changes in the conclusions will be brought forward into the main text.

14 Heterogeneity Analyses

14.1 By grade

The main outcome will be analyzed separately by grade and in pooled interaction models:

$$D_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \sum_g \theta_g Grade_g + \sum_g \phi_g (M_i \times Grade_g) + \sum_g \psi_g (A_i \times Grade_g) + \varepsilon_i. \tag{24}$$

14.2 By prior ability

If a pre-treatment measure of prior achievement is available:

$$D_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \eta Ability_i + \kappa_1 (M_i \times Ability_i) + \kappa_2 (A_i \times Ability_i) + \kappa_3 (M_i \times A_i \times Ability_i) + \delta_g + \varepsilon_i. \tag{25}$$

If only coarse bins are available, terciles or quartiles will be used.

14.3 By unpracticed-topic score

As an exploratory analysis, treatment effects will also be examined by unpracticed-topic score (for example, 0, 1, or 2 correct). Because this is a post-treatment measure, these analyses will be labeled exploratory and descriptive.

14.4 By topic

Where sample sizes allow, treatment effects may also be examined separately by topic bundle within grade.

14.5 By slowness exposure

As described above, exploratory heterogeneity analyses will examine whether treatment effects differ between students exposed to higher-slowness and lower-slowness periods.

15 Students Absent in Week 1 but Observed in Week 2

Students who miss the intervention session but appear in week 2 will not enter the main causal analysis. They may, however, be used in supplemental appendix tables as a descriptive non-exposed comparison group. These analyses will not be interpreted as randomized treatment effects.

16 Inference and Missing Data

16.1 Standard errors

Unless otherwise noted:

- student-level regressions will use heteroskedasticity-robust standard errors;
- question-level and prompt-level regressions will cluster standard errors at the student level.

If classroom-level implementation concerns are important and the number of classrooms is sufficient, classroom-clustered standard errors may be reported as a robustness check.

16.2 Multiple testing

Outcomes will be organized into the following families:

1. primary learning outcomes,
2. progression outcomes,
3. mastery-process outcomes,
4. AI-engagement outcomes,
5. robustness analyses related to slowness and low-effort testing,
6. heterogeneity and exploratory outcomes.

The combined practiced-minus-unpracticed delayed-test score is the single primary outcome. Exercise-specific delayed-test outcomes are key secondary outcomes. All other analyses are secondary or exploratory.

For the six planned primary contrasts, Holm-adjusted p-values will be reported alongside raw p-values. For the two exercise-specific delayed-test outcomes, raw and Holm-adjusted p-values will likewise be reported. Other secondary and exploratory families will be interpreted more cautiously, and false-discovery-rate adjusted q-values may be reported in appendix tables for larger families of related outcomes.

16.3 Missing data

The primary analysis will not impute missing delayed-test outcomes. Instead, the paper will:

- report missingness rates by treatment arm,
- test whether missingness is correlated with treatment,
- add sensitivity analyses if attrition is materially differential.

Possible sensitivity analyses include inverse-probability weighting and simple bounds.

17 Planned Tables

The tables below are formatted to resemble published working-paper tables: each has a clear title, a centered table body, and self-contained notes immediately below. The numeric values are illustrative placeholders only. They are included to make the intended presentation concrete and to clarify how the empirical patterns may ultimately be displayed.

Table 8A. NUMI features and motivation

Proposed title: *NUMI Support Features and Their Instructional Motivation*

Table 8A will provide a descriptive overview of the main features built into NUMI and the instructional motivation for including them in the intervention. The purpose of this table is to clarify what kinds of support were available to students assigned to AI, why those supports were designed as they were, and how they were expected to affect engagement and learning. This table complements Table 8, which reports how often students used these features, by explaining what the features actually do.

This table is descriptive only. It is intended to help the reader understand the design of the AI treatment rather than estimate causal effects.

NUMI Feature	Description	Instructional Motivation
Help get me started	Optional hint feature that gives students guidance on how to begin solving the current problem, typically without revealing the entire solution.	Designed to lower the barrier to starting a problem, especially when students are unsure how to translate the prompt into a first mathematical step. The feature is intended to promote productive struggle without leaving students stuck at the outset.
Solution walk through	Guided support sequence, sometimes triggered after an error, that walks students through the logic and steps needed to solve the problem.	Designed to help students recover from mistakes by slowing down the process and making the solution path explicit. Under mastery learning, this feature is especially important because students must resolve mistakes in order to continue.
Step solution explanation	Optional click-based explanation attached to a particular line or step of the worked solution.	Designed to provide targeted support exactly where the student is confused, rather than requiring them to read a full solution from the beginning. This feature aims to reduce friction by letting students request help only for the step they do not understand.
Open-ended interaction	Free-response text box that allows students to type a question, explanation, or short response to NUMI.	Designed to allow richer mathematical interaction than button-based prompts alone. This feature may support deeper reasoning and clarification, but it may also generate more friction because typing is slower and requires more effort.
Yes/No understanding checks	Low-friction button-based prompts asking whether a step or explanation makes sense.	Designed to quickly assess comprehension and route students to additional support when needed, while keeping interaction costs low.
A/B button-choice prompts	Multiple-choice style prompts embedded in the interaction, where students choose between two candidate answers or interpretations.	Designed to maintain engagement and test intermediate understanding in a structured way that is faster and less burdensome than typing.
Forced post-error interaction	NUMI support that appears automatically after some mistakes, requiring some engagement before the student can proceed.	Designed to ensure that at least some students engage with corrective support after an error, particularly in settings where they might otherwise skip ahead or disengage.

Notes: This table is descriptive and summarizes the main support features available in NUMI. The descriptions and motivations are intended to clarify how the AI treatment was designed to reduce friction, support recovery after mistakes, and encourage engagement with mathematical reasoning. The table does not report frequencies of use; those are reported separately in Table 8.

Table 8B. Illustrative sample NUMI conversations

Proposed title: *Illustrative Sample NUMI Conversations by Feature Type*

Table 8B will provide stylized examples of the kinds of exchanges students could have with NUMI. The purpose of this table is to make the AI treatment more concrete by showing what the interaction actually looked like from the student's perspective. These examples complement Tables 8 and 8A by illustrating the content and tone of the support features rather than their frequency or motivation.

This table is descriptive only. The examples are representative mock conversations, not verbatim transcripts.

Feature Type	Illustrative Sample Conversation
Help get me started	
textbfNUMI: Let's start with what the problem is asking. Which quantity do you need to find first?	
textbfStudent: I'm not sure.	
textbfNUMI: Try identifying the total and the number of groups. Once you have those, you can write the first equation.	
textbfStudent: Oh, so I should divide 24 by 3 first?	
textbfNUMI: Yes, that gives you the amount in each group. Now try entering that as your first step.	
addlinespace[0.5em] Solution walk through	
textbfNUMI: Let's work through this together. You chose an incorrect answer. What should we do first?	
textbfStudent: Multiply?	
textbfNUMI: Not yet. First, we need to isolate the variable. Which operation will undo the subtraction?	
textbfStudent: Add 5.	

Notes: This table presents stylized examples of NUMI interactions for four core feature types: Help get me started, Solution walk through, Step solution explanation, and open-ended interaction. The examples are illustrative and are included to help the reader understand the structure and tone of the AI support available during the intervention. They are not verbatim records of actual student conversations.

Table 1. Sample composition by grade, topic, mastery status, and AI status

Table 1: Sample Composition by Grade, Topic, Mastery Status, and AI Status

	Grade 6	Grade 7	Grade 8	Total
Total Number	1,240	1,315	1,180	3,735
Fraction Topic A	0.503	0.496	0.511	0.503
Fraction Topic B	0.497	0.504	0.489	0.497
Difference	0.006	-0.008	0.022	0.006
Fraction Mastery	0.418	0.447	0.465	0.443
Fraction Non-Mastery	0.582	0.553	0.535	0.557
Difference	-0.164***	-0.106***	-0.070**	-0.114***
Fraction AI	0.488	0.501	0.495	0.495
Fraction Non-AI	0.512	0.499	0.505	0.505
Difference	-0.024	0.002	-0.010	-0.010

Notes: This table describes the composition of the analysis sample by grade. Columns report values for Grade 6, Grade 7, Grade 8, and the full sample. The first row reports the total number of students in each column. Subsequent rows report the fraction of students assigned to Topic A and Topic B, the fraction classified as mastery and non-mastery, and the fraction classified as AI and non-AI. For each pair, the difference row reports the difference between the first-listed and second-listed category. Statistical significance for differences is indicated by asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2. Week-1 platform engagement and subsequent participation

Table 2: Week-1 Platform Engagement and Subsequent Participation by Grade, Topic, Mastery Status, and Speed Issues

Panel A. Total Sample

	N (Week 1 Login)	Started Video 1	Watched 75% of Video 1	Attempted Exercise 1	3 Correct on Ex. 1	Started Video 2	Watched 75% of Video 2	Attempted Exercise 2	3 Correct on Ex. 2	Submitted Ticket	Took Test 1 Week Later
Grade 6 Topic A	312	0.84	0.71	0.69	0.44	0.61	0.53	0.49	0.31	0.28	0.52
Grade 6 Topic B	308	0.82	0.69	0.67	0.42	0.59	0.50	0.47	0.29	0.27	0.50
Grade 7 Topic A	326	0.86	0.75	0.73	0.49	0.65	0.58	0.54	0.35	0.31	0.56
Grade 7 Topic B	327	0.85	0.73	0.72	0.47	0.63	0.56	0.52	0.33	0.30	0.55
Grade 8 Topic A	301	0.87	0.76	0.74	0.51	0.67	0.60	0.55	0.37	0.34	0.58
Grade 8 Topic B	299	0.85	0.74	0.72	0.48	0.64	0.57	0.53	0.34	0.31	0.56
Total	1,873	0.85	0.73	0.71	0.47	0.63	0.56	0.52	0.33	0.30	0.55
No speed issues	1,402	0.89	0.79	0.77	0.52	0.69	0.63	0.58	0.38	0.35	0.61
Slow Speed Issues	471	0.73	0.57	0.54	0.31	0.45	0.37	0.34	0.18	0.16	0.38

Panel B. Mastery Sample

	N (Week 1 Login)	Started Video 1	Watched 75% of Video 1	Attempted Exercise 1	3 Correct on Ex. 1	Started Video 2	Watched 75% of Video 2	Attempted Exercise 2	3 Correct on Ex. 2	Submitted Ticket	Took Test 1 Week Later
Grade 6 Topic A	131	0.88	0.79	0.76	0.59	0.71	0.64	0.60	0.44	0.39	0.62
Grade 6 Topic B	128	0.87	0.78	0.75	0.57	0.69	0.63	0.58	0.42	0.38	0.61
Grade 7 Topic A	147	0.90	0.82	0.80	0.63	0.74	0.68	0.64	0.47	0.42	0.66
Grade 7 Topic B	147	0.89	0.81	0.79	0.61	0.73	0.66	0.62	0.45	0.41	0.65
Grade 8 Topic A	141	0.91	0.84	0.82	0.66	0.77	0.72	0.68	0.51	0.46	0.69
Grade 8 Topic B	139	0.90	0.82	0.80	0.63	0.75	0.69	0.65	0.48	0.43	0.67
Total	833	0.89	0.81	0.79	0.61	0.73	0.67	0.63	0.46	0.42	0.65
No speed issues	651	0.92	0.85	0.83	0.65	0.78	0.73	0.69	0.52	0.47	0.70
Slow Speed Issues	182	0.80	0.67	0.64	0.45	0.55	0.46	0.42	0.24	0.22	0.48

Panel C. Non-Mastery Sample

	N (Week 1 Login)	Started Video 1	Watched 75% of Video 1	Attempted Exercise 1	3 Correct on Ex. 1	Started Video 2	Watched 75% of Video 2	Attempted Exercise 2	3 Correct on Ex. 2	Submitted Ticket	Took Test 1 Week Later
Grade 6 Topic A	181	0.81	0.66	0.63	0.33	0.54	0.45	0.41	0.22	0.20	0.44
Grade 6 Topic B	180	0.78	0.63	0.61	0.31	0.51	0.42	0.39	0.20	0.19	0.42
Grade 7 Topic A	179	0.82	0.69	0.67	0.37	0.57	0.49	0.46	0.25	0.22	0.47
Grade 7 Topic B	180	0.82	0.67	0.66	0.36	0.55	0.47	0.44	0.24	0.21	0.46
Grade 8 Topic A	160	0.84	0.69	0.67	0.38	0.58	0.49	0.45	0.26	0.23	0.48
Grade 8 Topic B	160	0.81	0.67	0.65	0.35	0.54	0.46	0.43	0.22	0.20	0.46
Total	1,040	0.81	0.67	0.65	0.35	0.55	0.46	0.43	0.23	0.21	0.45
No speed issues	751	0.86	0.73	0.71	0.42	0.62	0.53	0.49	0.29	0.27	0.53
Slow Speed Issues	289	0.68	0.50	0.49	0.18	0.36	0.27	0.25	0.08	0.07	0.23

Notes: This table reports student progression through the platform during the first week and the immediate follow-up period. Each panel shows the number of students who logged into the platform in week 1 and the fraction who reached each subsequent stage: started the first video, watched at least 75 percent of the first video, attempted the first exercise, answered three questions in a row correctly on the first exercise, started the second video, watched at least 75 percent of the second video, attempted the second exercise, answered three questions in a row correctly on the second exercise, submitted an exercise ticket, and took the delayed test one week later. Rows are reported separately for Grade 6 Topic A, Grade 6 Topic B, Grade 7 Topic A, Grade 7 Topic B, Grade 8 Topic A, Grade 8 Topic B, and the total sample. Each panel also reports summary rows for students with no speed issues and students with slow speed issues. Fractions are measured within the row group shown.

Table 3. Early participation, mistakes, and follow-up completion

Proposed title: *Early Participation, Mistakes, Exit-Ticket Submission, and Delayed-Test Take-Up*

Table 3 will report treatment differences in four key process outcomes that track whether students engage with the platform and persist through the activity. The columns will report: (1) an indicator for attempting at least one question, (2) an indicator for making at least one mistake on any question, (3) an indicator for submitting the exit ticket in week 1, and (4) an indicator for taking the delayed test in week 2. These outcomes are useful for understanding both initial participation and later follow-through.

The estimating equation for each column is:

$$Y_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i, \tag{26}$$

where Y_i is the process outcome of interest, M_i is an indicator for mastery assignment, A_i is an indicator for AI assignment, and δ_g are grade fixed effects.

The table would look as follows:

	Attempted at Least One Question	Made at Least One Mistake	Submitted Exit Ticket in Week 1	Took Test in Week 2
Mastery	-0.012	0.038**	-0.136***	-0.021
	(0.011)	(0.018)	(0.021)	(0.017)
AI	0.009	0.026	0.012	0.014
	(0.011)	(0.017)	(0.020)	(0.017)
Mastery × AI	0.017	0.029	0.041*	0.022
	(0.015)	(0.024)	(0.025)	(0.023)
Control Mean	0.781	0.462	0.341	0.889
Observations	3,735	3,735	3,735	3,960
Grade FE	Yes	Yes	Yes	Yes

Notes: Each column reports a separate regression of the indicated process outcome on mastery assignment, AI assignment, and their interaction, with grade fixed effects included. Attempted at least one question is an indicator equal to one if the student attempted any question on the platform during week 1. Made at least one mistake is an indicator equal to one if the student answered at least one attempted question incorrectly. Submitted exit ticket in week 1 is an indicator equal to one if the student submitted the exit ticket during the intervention session. Took test in week 2 is an indicator equal to one if the student was observed taking the delayed test one week later. The control mean reports the mean of the dependent variable in the omitted category, Non-mastery + no AI. Standard errors are shown in parentheses.

Table 4. Main delayed-test learning effects

Proposed title: *Main Delayed-Test Learning Effects*

Table 4 is the core outcome table. It reports treatment effects on the practiced-minus-unpracticed delayed-test score, separately for the combined outcome and the two exercise-specific outcomes. It also reports the practiced-topic score and unpracticed-topic score as secondary outcomes. The constant is included because it is substantively important: it gives the practiced-minus-unpracticed difference for students in the omitted category, Non-mastery + no AI. That is, it captures the effect of receiving the treated topic relative to the non-treated topic in the baseline treatment arm.

The estimating equation is:

$$D_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i, \tag{27}$$

with the same form applied to each outcome shown in the columns.

The table would look as follows:

	(1) Combined	(2) Ex. 1	(3) Ex. 2	(4) Practiced	(5) Unpracticed
Constant	0.182***	0.101***	0.081***	0.944***	0.762***
	(0.028)	(0.021)	(0.020)	(0.054)	(0.049)
Mastery	0.082**	0.109***	-0.027	0.064*	-0.018
	(0.033)	(0.036)	(0.032)	(0.037)	(0.028)
AI	0.046	0.031	0.015	0.053	0.007
	(0.032)	(0.035)	(0.031)	(0.036)	(0.027)
Mastery × AI	0.094**	0.071*	0.023	0.118**	0.024
	(0.041)	(0.043)	(0.039)	(0.048)	(0.034)
Grade FE	Yes	Yes	Yes	Yes	Yes
Observations	3,418	3,418	3,418	3,418	3,418
Mean dep. var.	0.214	0.127	0.087	1.043	0.829

Notes: Each column reports a separate regression of the indicated delayed-test outcome on mastery assignment, AI assignment, and their interaction, with grade fixed effects included. Column 1 reports the combined practiced-minus-unpracticed score. Columns 2 and 3 report the same difference separately for the exercise 1 and exercise 2 items. Columns 4 and 5 report the total practiced-topic and unpracticed-topic scores. The constant gives the estimated value of the dependent variable for students in the omitted category, Non-mastery + no AI. In Columns 1–3, this can be interpreted as the estimated effect of receiving the treated topic relative to the non-treated topic in the baseline arm. Standard errors are shown in parentheses.

Table 4A. Main delayed-test learning effects by grade

Proposed title: *Main Delayed-Test Learning Effects by Grade*

This companion table reports the same core delayed-test learning specification separately by grade. The purpose is to show how the baseline practiced-minus-unpracticed difference and the treatment modifications to that difference vary across Grade 6, Grade 7, and Grade 8.

The estimating equation within each grade is:

$$D_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \varepsilon_i, \tag{28}$$

where the constant again gives the practiced-minus-unpracticed difference for Non-mastery + no AI students within that grade.

The table would look as follows:

	Grade 6	Grade 7	Grade 8
Constant	0.156***	0.181***	0.209***
	(0.046)	(0.043)	(0.042)
Mastery	0.054	0.083*	0.108**
	(0.051)	(0.046)	(0.045)
AI	0.021	0.039	0.072*
	(0.049)	(0.044)	(0.043)
Mastery × AI	0.072	0.095*	0.111**
	(0.064)	(0.056)	(0.054)
Observations	1,128	1,171	1,119
Mean dep. var.	0.187	0.214	0.241

Notes: Each column reports the primary delayed-test regression separately by grade, using the combined practiced-minus-unpracticed delayed-test score as the dependent variable. The constant gives the estimated practiced-minus-unpracticed difference for students in the omitted category, Non-mastery + no AI, within that grade. Standard errors are shown in parentheses.

Table 5. Pairwise treatment comparisons

Proposed title: *Pairwise Treatment Comparisons from the Primary Learning Regression*

Table 5 will translate the coefficients from the main delayed-test regression into substantively meaningful treatment comparisons. Rather than repeating the regression coefficients, this table will report the linear combinations corresponding to the AI effect among non-mastery students, the AI effect among mastery students, the mastery effect among no-AI students, the mastery effect among AI students, the difference between Mastery + AI and Non-mastery + no AI, and the complementarity term.

The estimating equation underlying these comparisons is:

$$D_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i, \tag{29}$$

where D_i is the practiced-minus-unpracticed delayed-test score, M_i is an indicator for mastery assignment, A_i is an indicator for AI assignment, and δ_g are grade fixed effects.

The reported pairwise comparisons correspond to the following linear combinations:

$$\begin{aligned}
\text{AI effect among non-mastery students:} &\quad \beta_2, \\
\text{AI effect among mastery students:} &\quad \beta_2 + \beta_3, \\
\text{Mastery effect among no-AI students:} &\quad \beta_1, \\
\text{Mastery effect among AI students:} &\quad \beta_1 + \beta_3, \\
\text{Mastery + AI vs. Non-mastery + no AI:} &\quad \beta_1 + \beta_2 + \beta_3, \\
\text{Complementarity:} &\quad \beta_3.
\end{aligned}$$

The table would look as follows:

Comparison	Estimate	Standard Error
AI effect among non-mastery students	0.046	(0.032)
AI effect among mastery students	0.140**	(0.050)
Mastery effect among no-AI students	0.082**	(0.033)
Mastery effect among AI students	0.176***	(0.046)
Mastery + AI vs. Non-mastery + no AI	0.222***	(0.044)
Complementarity	0.094**	(0.041)

Notes: This table reports linear combinations from the primary delayed-test regression in Table 4. The AI effect among mastery students is $\beta_2+\beta_3$, the mastery effect among AI students is $\beta_1+\beta_3$, the difference between Mastery + AI and Non-mastery + no AI is $\beta_1+\beta_2+\beta_3$, and complementarity is β_3. Standard errors are obtained using the estimated variance-covariance matrix from the primary regression. Statistical significance is indicated by asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6. Progression outcomes across all students

Proposed title: *Progression Outcomes Across All Students*

Table 6 will summarize how mastery and AI affect students' movement through the activity. These are the key breadth-of-completion outcomes, and they help interpret the learning results in Table 4. In particular, the table is meant to show whether mastery increases depth of engagement at the cost of reducing progression to later material, and whether AI offsets any of that reduction by helping students move through the activity more successfully. The outcomes therefore track whether students reach later stages of the sequence, get to the exit ticket, and complete more work overall.

The estimating equation for each column is:

$$Y_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i, \tag{30}$$

where Y_i is the progression outcome of interest, M_i is an indicator for mastery assignment, A_i is an indicator for AI assignment, and δ_g are grade fixed effects.

The table would look as follows:

	Reached Video 2	Attempted Exercise 2	Completed Set 2	Reached Exit Ticket	Submitted Exit Ticket	Total Questions Attempted	Total Time (Minutes)
Constant	0.662***	0.528***	0.392***	0.341***	0.328***	8.14***	26.8***
	(0.024)	(0.024)	(0.023)	(0.022)	(0.022)	(0.41)	(0.72)
Mastery	-0.118***	-0.127***	-0.132***	-0.141***	-0.136***	-1.84***	3.12***
	(0.018)	(0.019)	(0.020)	(0.021)	(0.021)	(0.42)	(0.74)
AI	0.021	0.018	0.014	0.009	0.012	0.63	1.48**
	(0.017)	(0.018)	(0.018)	(0.019)	(0.019)	(0.40)	(0.69)
Mastery × AI	0.044**	0.051**	0.048**	0.039*	0.041*	0.95*	0.88
	(0.022)	(0.023)	(0.024)	(0.024)	(0.024)	(0.50)	(0.83)
Observations	3,735	3,735	3,735	3,735	3,735	3,735	3,735
Mean dep. var.	0.63	0.52	0.33	0.31	0.30	7.84	28.6
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each column reports a separate regression of the indicated progression outcome on mastery assignment, AI assignment, and their interaction, with grade fixed effects included. The constant gives the estimated value of the dependent variable for students in the omitted category, Non-mastery + no AI. Binary outcomes are coded as indicators. Total questions attempted and total time on platform are continuous outcomes. Standard errors are shown in parentheses. Statistical significance is indicated by asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7. Set completion and progression requirements by mastery status

Proposed title: *Set Completion, Progression Requirements, and Time to Progress by Mastery Status*

Table 7 will examine how AI affects students' ability to satisfy the progression requirements of the activity, separately within the mastery and non-mastery conditions. The purpose of this table is to make the comparison between the two regimes more concrete. Under mastery, students must watch at least one minute of the video and answer three questions in a row correctly in order to move forward. Under non-mastery, students can move more freely, so analogous outcomes capture whether they completed the corresponding set and how much effort or time was required before doing so. This table therefore helps show whether AI changes students' ability to complete the first and second instructional sets, how many attempts it takes, and how long progress takes under the two assignment regimes.

In Panel A, the sample is restricted to students assigned to mastery. In Panel B, the sample is restricted to students assigned to non-mastery. The coefficients therefore show the effect of AI within each regime.

The estimating equation for each column within each panel is:

$$Y_i = \alpha + \beta A_i + \delta_g + \varepsilon_i, \tag{31}$$

where Y_i is the process outcome of interest, A_i is an indicator for AI assignment, and δ_g are grade fixed effects.

The table would look as follows:

	Completed First Set	Completed Second Set	Completed Both Sets	Attempts to First Set	Time to First Set	Attempts to Second Set	Time to Second Set
Panel A. Mastery Sample							
AI	0.063**	0.051**	0.047**	-0.42*	-1.18	-0.37	-0.94
	(0.025)	(0.024)	(0.023)	(0.24)	(0.81)	(0.31)	(1.04)
Constant	0.71***	0.46***	0.43***	4.28***	9.84***	3.91***	8.62***
	(0.03)	(0.03)	(0.03)	(0.29)	(0.92)	(0.36)	(1.15)
Observations	1,789	1,789	1,789	1,789	1,789	1,204	1,204
Mean dep. var.	0.74	0.49	0.46	4.11	9.43	3.78	8.41
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Panel B. Non-Mastery Sample							
AI	0.018	0.014	0.012	-0.11	-0.36	-0.09	-0.28
	(0.019)	(0.018)	(0.018)	(0.17)	(0.54)	(0.21)	(0.67)
Constant	0.79***	0.54***	0.51***	2.67***	6.21***	2.39***	5.74***
	(0.02)	(0.02)	(0.02)	(0.20)	(0.61)	(0.24)	(0.75)
Observations	1,946	1,946	1,946	1,946	1,946	1,052	1,052
Mean dep. var.	0.80	0.55	0.52	2.61	6.09	2.34	5.66
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports regressions estimated separately within the mastery and non-mastery samples. In Panel A, *Completed First Set* is an indicator equal to one if the student satisfied the first mastery requirement and advanced past the first video-exercise set; *Completed Second Set* is an indicator equal to one if the student satisfied the second mastery requirement and advanced past the second set; and *Completed Both Sets* is an indicator equal to one if the student completed both required sets. *Attempts to First Set* and *Attempts to Second Set* record the number of exercise attempts made before completing the corresponding set, and *Time to First Set* and *Time to Second Set* record elapsed time in minutes before completing the corresponding set. In Panel B, the analogous outcomes are defined using completion of the first and second instructional sets under the non-mastery rules, where progression does not require three consecutive correct answers. The constant gives the mean outcome for students in the omitted category within each panel, namely no-AI students. Standard errors are shown in parentheses. Statistical significance is indicated by asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8. AI engagement outcomes

Proposed title: *AI Engagement Outcomes Among AI-Assigned Students*

Table 8 will examine whether mastery increases students' take-up and intensity of AI use among those assigned access to NUMI. This table is central to the mechanism analysis because the main hypothesis of the paper is not simply that AI helps on its own, but that AI may be more effective when mastery learning gives students stronger incentives to resolve mistakes rather than move on. The outcomes therefore track whether students use NUMI at all, how often they interact with it, how much time they spend with it, and which features they use.

Because all students in this table are assigned to AI, the coefficient of interest is the effect of mastery assignment on AI engagement.

The estimating equation for each column is:

$$Y_i = \alpha + \beta M_i + \delta_g + \varepsilon_i, \tag{32}$$

where Y_i is the AI-engagement outcome of interest, M_i is an indicator for mastery assignment, and δ_g are grade fixed effects.

The table would look as follows:

	Any NUMI Use	NUMI Interactions	Total AI Time (Min.)	Any Hint Use	Any Typed Message	Forced WT Completed	Clicked Step Explanation
Mastery	0.081***	1.94***	2.36***	0.067***	0.029*	0.118***	0.041**
	(0.021)	(0.42)	(0.58)	(0.019)	(0.017)	(0.024)	(0.018)
Constant	0.50***	3.88***	5.24***	0.29***	0.16***	0.19***	0.18***
	(0.02)	(0.31)	(0.44)	(0.02)	(0.02)	(0.02)	(0.02)
Observations	1,867	1,867	1,867	1,867	1,867	1,867	1,867
Mean dep. var.	0.54	4.82	6.41	0.32	0.18	0.27	0.21
Grade FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table restricts the sample to students assigned to AI. Each column reports a separate regression of the indicated AI-engagement outcome on mastery assignment, with grade fixed effects included. *Any NUMI Use* is an indicator equal to one if the student interacted with NUMI at least once. *NUMI Interactions* counts the total number of logged NUMI interactions. *Total AI Time* measures total minutes spent interacting with NUMI. *Any Hint Use* is an indicator for whether the student used the "help get me started" feature at least once. *Any Typed Message* is an indicator for whether the student entered any open-ended typed response. *Forced WT Completed* is an indicator for whether the student completed at least one forced walkthrough after making a mistake. *Clicked Step Explanation* is an indicator for whether the student clicked for at least one optional step explanation. The constant gives the mean outcome for students in the omitted category, non-mastery students assigned to AI. Standard errors are shown in parentheses. Statistical significance is indicated by asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9. Prompt-type response patterns

Table 3: Prompt-Type Response Patterns and Friction

Prompt Type	Response Rate	Avg. Latency (Sec.)	Nonresponse	Continued to Next Prompt	Continued to
Yes/No prompt	0.91	4.8	0.09	0.88	0.8
A/B button prompt	0.94	3.9	0.06	0.90	0.8
Open-ended prompt	0.63	11.7	0.37	0.71	0.6

Notes: This table summarizes prompt-level outcomes by prompt type among AI-assigned students. Values are descriptive means ra estimates. Response rate is the share of prompts that receive a response. Continued to next prompt and continued to next problem a whether the student remains engaged immediately following the prompt.

Table 10. Heterogeneity by grade

Table 4: Heterogeneity in the Main Learning Effect by Grade

	Grade 6	Grade 7	Grade 8	Pooled Interaction Test
Mastery	0.054	0.083*	0.108**	$p = 0.214$
	(0.051)	(0.046)	(0.045)	
AI	0.021	0.039	0.072*	$p = 0.327$
	(0.049)	(0.044)	(0.043)	
Mastery × AI	0.072	0.095*	0.111**	$p = 0.441$
	(0.064)	(0.056)	(0.054)	
Observations	1,128	1,171	1,119	3,418

Notes: Columns 1–3 report separate regressions of the main practiced-minus-unpracticed delayed-test score by grade. The final column reports p-values from pooled interaction tests of whether the corresponding treatment effect differs across grades. Standard errors are shown in parentheses in the grade-specific columns.

Table 11. Heterogeneity by prior ability

Table 5: Heterogeneity in the Main Learning Effect by Prior Ability

	Main Outcome	Ex. 1 Outcome	Ex. 2 Outcome
Mastery	0.079**	0.102**	-0.023
	(0.035)	(0.039)	(0.034)
AI	0.043	0.028	0.015
	(0.034)	(0.037)	(0.033)
Mastery × AI	0.087**	0.068	0.019
	(0.042)	(0.046)	(0.040)
Prior ability (SD)	0.118***	0.071***	0.047**
	(0.018)	(0.021)	(0.019)
Mastery × Prior ability	-0.018	-0.009	-0.009
	(0.021)	(0.024)	(0.022)
AI × Prior ability	0.026	0.011	0.015
	(0.020)	(0.023)	(0.021)
Mastery × AI × Prior ability	0.041*	0.037	0.004
	(0.024)	(0.027)	(0.025)

Notes: Each column reports a regression that interacts treatment assignment with a standardized prior-ability measure. Coefficients on the interaction terms indicate whether treatment effects differ systematically with prior achievement. Standard errors are shown in parentheses. Statistical significance is indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 12. Exploratory heterogeneity by unpracticed-topic score

Proposed title: *Exploratory Heterogeneity in Learning Effects by Unpracticed-Topic Score*

Table 12 will examine whether the main delayed-test treatment effects differ across students with different levels of performance on the unpracticed-topic questions. The motivation for this table is that the unpracticed-topic score may proxy for general math understanding, effort on the delayed test, or broader preparedness on material not directly practiced during the intervention. The table is therefore useful for describing whether the treatment effects appear larger or smaller among students who perform worse or better on the unpracticed material.

Because the unpracticed-topic score is itself measured after treatment, this table is explicitly exploratory and descriptive rather than causal. It should be interpreted as a way to organize the results, not as a clean test of pre-treatment heterogeneity.

The estimating equation within each subgroup is:

$$D_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i, \tag{33}$$

where D_i is the practiced-minus-unpracticed delayed-test score, M_i is an indicator for mastery assignment, A_i is an indicator for AI assignment, and δ_g are grade fixed effects. The regression is estimated separately for students with unpracticed-topic score equal to 0, 1, or 2.

The table would look as follows:

	Unpracticed Score = 0	Unpracticed Score = 1	Unpracticed Score = 2
Constant	0.144***	0.176***	0.205***
	(0.041)	(0.037)	(0.044)
Mastery	0.061	0.084*	0.097*
	(0.048)	(0.045)	(0.053)
AI	0.033	0.041	0.056
	(0.046)	(0.043)	(0.051)
Mastery × AI	0.101*	0.089*	0.072
	(0.058)	(0.052)	(0.061)
Observations	924	1,481	1,013
Mean dep. var.	0.173	0.214	0.249
Grade FE	Yes	Yes	Yes

Notes: This table is exploratory because the heterogeneity grouping variable is post-treatment. Each column reports the primary delayed-test regression separately for students in the indicated unpracticed-topic score group. The constant gives the estimated practiced-minus-unpracticed delayed-test difference for students in the omitted category, Non-mastery + no AI, within that subgroup. Standard errors are shown in parentheses. Statistical significance is indicated by asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 13. Supplemental descriptive analysis including week-2-only students

Table 6: Supplemental Descriptive Comparisons Including Week-2-Only Students

	Main Analysis Sample	Week-2-Only Students
Grade 6 indicator	0.330	0.357
Female indicator	0.489	0.476
Prior achievement (SD)	0.014	-0.221
Delayed-test total score	1.87	1.21
Answered all delayed-test items	0.91	0.74
Completed delayed test in under 3 min	0.06	0.18
Observations	3,418	84

Notes: This table is descriptive only. Week-2-only students were not exposed to the intended practice treatment and are therefore excluded from the main causal analysis. The purpose is to characterize their composition and outcome quality relative to the main analysis sample.

Table 14. Platform slowness and treatment-effect heterogeneity

Proposed title: *Platform Slowness During the Intervention and Main Learning Results Across Slowness-Restricted Samples*

Table 14 will summarize platform slowness during the intervention and then examine whether the main delayed-test learning results are stronger in samples exposed to better platform performance. The motivation for this table is that website slowness may have directly reduced the quality of the intervention, especially for students assigned to AI, whose treatment experience depended more heavily on interactive responsiveness. The first part of the table is descriptive and is intended to show how platform performance varied across the four experimental days and between morning and afternoon sessions. The second part reports the main learning regression on progressively more restrictive samples defined by student-level slowness exposure.

The table is organized in two panels. Panel A presents descriptive slowness measures by day and time of day. Panel B reports the main delayed-test learning regression separately for samples defined by alternative slowness cutoffs: students whose average wait time was less than 2 seconds, less than 5 seconds, less than 10 seconds, greater than 10 seconds, and students observed on Days 3–4 only. The purpose is to show whether the main treatment patterns are attenuated in slower platform environments and whether they become stronger in the cleaner implementation periods.

The estimating equation in Panel B is:

$$D_i = \alpha + \beta_1 M_i + \beta_2 A_i + \beta_3 (M_i \times A_i) + \delta_g + \varepsilon_i, \tag{34}$$

where D_i is the practiced-minus-unpracticed delayed-test score, M_i is an indicator for mastery assignment, A_i is an indicator for AI assignment, and δ_g are grade fixed effects.

The table would look as follows:

	Avg. Wait (Sec.)	Median Wait (Sec.)	Share Waits > 10 Sec.	Share Waits > 15 Sec.	Students Observed
Panel A. Descriptive Slowness by Day and Time of Day					
Day 1 Morning	10.9	8.4	0.27	0.13	486
Day 1 Afternoon	14.8	11.2	0.40	0.23	504
Day 2 Morning	9.1	7.0	0.22	0.10	492
Day 2 Afternoon	12.1	9.3	0.31	0.17	498
Day 3 Morning	5.1	3.9	0.08	0.03	487
Day 3 Afternoon	6.8	5.2	0.14	0.06	503
Day 4 Morning	4.2	3.1	0.06	0.02	479
Day 4 Afternoon	5.4	4.0	0.10	0.04	511
Overall	8.5	6.2	0.20	0.10	3,960

Panel B. Main Learning Results in Samples Defined by Slowness Exposure

	Avg. Wait < 2 Sec.	Avg. Wait < 5 Sec.	Avg. Wait < 10 Sec.	Avg. Wait > 10 Sec.	Days 3–4 Only
Constant	0.214***	0.198***	0.186***	0.143**	0.205***
	(0.061)	(0.041)	(0.032)	(0.057)	(0.039)
Mastery	0.118**	0.101**	0.089**	0.024	0.096**
	(0.057)	(0.045)	(0.039)	(0.062)	(0.047)
AI	0.079	0.063*	0.054	0.008	0.071*
	(0.054)	(0.038)	(0.035)	(0.059)	(0.041)
Mastery × AI	0.129**	0.118**	0.106**	0.031	0.121**
	(0.061)	(0.051)	(0.046)	(0.071)	(0.053)
Observations	624	1,744	2,981	437	1,711
Mean dep. var.	0.263	0.236	0.221	0.118	0.248
Grade FE	Yes	Yes	Yes	Yes	Yes

Notes: Panel A reports descriptive measures of platform slowness separately by experimental day and by morning versus afternoon sessions. Average and median wait times are measured as the number of seconds between a student request and the next platform response. The two share variables report the fraction of waits exceeding 10 seconds and 15 seconds, respectively. Panel B reports the main practiced-minus-unpracticed delayed-test learning regression estimated separately on samples defined by student-level slowness exposure. Columns 1–4 restrict the sample using the student's average wait time during the intervention session. Column 5 restricts the sample to students observed on experimental Days 3–4 only, when platform performance had improved. The constant gives the practiced-minus-unpracticed delayed-test difference for students in the omitted category, Non-mastery + no AI, within each restricted sample. These analyses are exploratory and are intended to assess whether the main learning results are stronger in better-functioning platform environments. Standard errors are shown in parentheses. Statistical significance is indicated by asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 15. Robustness to excluding students exposed to slow periods

Table 7: Robustness to Excluding Students Exposed to Slow Periods

	Full Sample	Excl. Slow 1	Excl. Slow 2	Excl. Days 1–2
Mastery	0.082**	0.089**	0.094**	0.101**
	(0.033)	(0.035)	(0.037)	(0.041)
AI	0.046	0.054	0.061*	0.068*
	(0.032)	(0.033)	(0.035)	(0.039)
Mastery × AI	0.094**	0.106**	0.113**	0.127**
	(0.041)	(0.043)	(0.046)	(0.051)
Grade FE	Yes	Yes	Yes	Yes
Observations	3,418	3,071	2,824	1,711

Notes: Each column reports the primary delayed-test regression on a different sample restriction. The restricted samples exclude students based on alternative measures of platform slowness. These analyses are exploratory and intended to assess sensitivity to the intervention environment. Standard errors are shown in parentheses. Statistical significance is indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 16. Low-effort delayed-test completion

Table 8: Low-Effort Delayed-Test Completion and Robustness to Exclusions

	Full Sample	Excl. 0 Answers	Excl. < 3 Min.
Panel A. Main Outcome Robustness			
Mastery	0.082**	0.086**	0.091**
	(0.033)	(0.034)	(0.035)
AI	0.046	0.049	0.052
	(0.032)	(0.032)	(0.033)
Mastery × AI	0.094**	0.101**	0.108**
	(0.041)	(0.042)	(0.043)
Observations	3,418	3,302	3,244
Panel B. Low-Effort Incidence by Treatment			
Mastery	0.006		
	(0.008)		
AI	0.004		
	(0.008)		
Mastery × AI	-0.003		
	(0.011)		
Control Mean	0.071		

Notes: Panel A reports the main delayed-test regression under alternative exclusions for low-effort delayed-test completion. Panel B reports a regression of the indicator for low-effort delayed-test behavior on mastery assignment, AI assignment, and their interaction. Candidate low-effort definitions include answering no delayed-test questions and completing the delayed test in less than three minutes. These analyses are exploratory. Standard errors are shown in parentheses. Statistical significance is indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

18 Planned Figures

The figures below are formatted to resemble published working-paper figures. Each figure has a title and a note placed immediately below. Because this is still a mock pre-analysis draft, the figure panels are shown as schematic placeholders that mimic the final visual layout.

Figure 8. Platform slowness over time

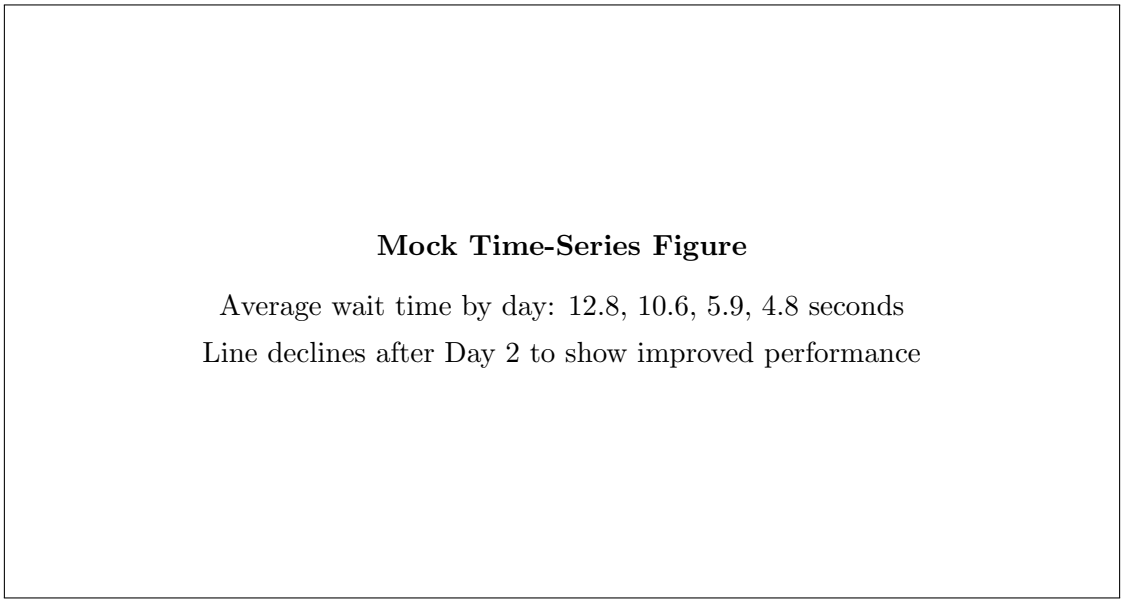

Figure 1: Platform Slowness Over the Experimental Period

Notes: The final figure will plot average wait times by experimental day, class period, or clock time and highlight the improvement in platform performance after the first two days.

Variables Needed in a Single Spreadsheet

Below is a practical list of variables to include in one student-level analysis spreadsheet. The goal is to keep this list as simple as possible while still covering the outcomes and specifications used in the current pre-analysis plan. In particular, the list focuses on treatment assignment, week-1 platform behavior, delayed-test outcomes, AI usage, and platform slowness.

A. Student identifiers and treatment assignment

- `student_id`: unique student identifier
- `grade`: grade level
- `ai_assign`: 1 if assigned AI, 0 otherwise
- `mastery_assign`: 1 if assigned mastery, 0 otherwise
- `topic_assign`: Topic A / Topic B indicator
- `treatment_arm`: combined four-arm treatment label

B. Week-1 participation and progression

These variables are used for the platform-engagement, progression, and process tables.

- `logged_in_week1`: logged into platform in week 1
- `started_video1`: started first video
- `watched_video1_75`: watched at least 75 percent of first video
- `attempted_exercise1`: attempted first exercise
- `num_attempted_exercise1`: number of questions attempted from first exercise
- `num_correct_exercise1`: number of questions correct from first exercise
- `three_correct_exercise1`: answered three questions in a row correctly on first exercise
- `started_video2`: started second video
- `watched_video2_75`: watched at least 75 percent of second video
- `attempted_exercise2`: attempted second exercise
- `num_attempted_exercise2`: number of questions attempted from second exercise
- `num_correct_exercise2`: number of questions correct from second exercise
- `three_correct_exercise2`: answered three questions in a row correctly on second exercise

- `attempted_any_question`: attempted at least one question in week 1
- `made_any_mistake`: made at least one mistake on any question
- `completed_set1`: completed the first instructional set
- `completed_set2`: completed the second instructional set
- `reached_exit_ticket`: reached the exit ticket
- `submitted_exit_ticket`: submitted the exit ticket
- `exit_ticket_score`: exit ticket score
- `total_questions_attempted`: total number of exercise questions attempted
- `total_platform_time_sec`: total time on platform in seconds

C. Mastery-specific process variables

These are needed for the within-mastery process analysis.

- `cleared_gate1`: satisfied the first mastery requirement
- `cleared_gate2`: satisfied the second mastery requirement
- `cleared_both_gates`: satisfied both mastery requirements
- `attempts_to_gate1`: number of attempts before clearing the first gate
- `attempts_to_gate2`: number of attempts before clearing the second gate
- `time_to_gate1_sec`: elapsed time before clearing the first gate
- `time_to_gate2_sec`: elapsed time before clearing the second gate

D. Delayed-test outcomes

These are the essential variables for the main estimands.

- `took_test_week2`: 1 if the student took the delayed test in week 2
- `y_p1`: score on the delayed-test question corresponding to practiced exercise 1
- `y_p2`: score on the delayed-test question corresponding to practiced exercise 2
- `y_np1`: score on the delayed-test question corresponding to unpracticed exercise 1
- `y_np2`: score on the delayed-test question corresponding to unpracticed exercise 2

- `delayed_test_total`: total delayed-test score
- `delayed_test_answered_count`: number of delayed-test questions answered
- `delayed_test_duration_sec`: total delayed-test duration in seconds

You can then pre-compute:

- `d_combined` $= (y\_p1 + y\_p2) - (y\_np1 + y\_np2)$
- `d_ex1` $= y\_p1 - y\_np1$
- `d_ex2` $= y\_p2 - y\_np2$
- `practiced_score` $= y\_p1 + y\_p2$
- `unpracticed_score` $= y\_np1 + y\_np2$

E. Delayed-test low-effort indicators

- `loweffort_zeroanswers`: 1 if answered 0 delayed-test questions
- `loweffort_under3min`: 1 if `delayed_test_duration_sec` < 180
- `loweffort_any`: 1 if either of the above

F. Error-recovery outcomes

These are only needed if the exploratory post-error analyses are retained.

- `first_mistake_occurred`: indicator
- `eventual_success_after_firstmistake`
- `gate_completion_after_firstmistake`
- `attempts_from_firstmistake_to_nextcorrect`
- `time_from_firstmistake_to_nextcorrect_sec`
- `next_attempt_correct_after_firstmistake`

G. AI usage outcomes

These are the student-level variables needed for the AI engagement table.

- numi_any_use time_to_first_numi_use_sec: elapsed time from login to first NUMI use
- numi_use_before_firstmistake: 1 if the student uses NUMI before the first recorded mistake
- numi_use_after_firstmistake: 1 if the student uses NUMI after the first recorded mistake
- numi_interaction_count
- numi_interactions_per_attempted_question:
- numi_total_time_sec
- hint_any_use
- hint_count
- typed_message_any
- typed_message_count
- forced_walkthrough_count
- forced_walkthrough_completed_count
- clicked_step_explanation_count
- button_interaction_share

H. Prompt-type aggregates at the student level

These are only needed if the prompt-type table is retained.

- yesno_prompt_count
- ab_prompt_count
- openended_prompt_count
- yesno_response_rate
- ab_response_rate
- openended_response_rate
- yesno_avg_latency_sec
- ab_avg_latency_sec
- openended_avg_latency_sec

I. Platform slowness variables

These are needed for the appendix slowness analyses and the slowness-restricted samples.

- `experiment_day`: day 1, 2, 3, or 4
- `class_period`: class period or time block
- `avg_wait_sec`
- `median_wait_sec`
- `share_waits_over10`
- `max_wait_sec`: optional
- `request_count`: optional
- `peer_avg_wait_sec_leaveout`: leave-one-out average wait time among other students active in the same class period or narrow time window, if recoverable from the logs

You can then create the restricted-sample indicators used in the slowness tables:

- `avgwait_lt2`: 1 if average wait time is less than 2 seconds
- `avgwait_lt5`: 1 if average wait time is less than 5 seconds
- `avgwait_lt10`: 1 if average wait time is less than 10 seconds
- `avgwait_gt10`: 1 if average wait time is greater than 10 seconds
- `peer_slow_period`: 1 if the leave-one-out contemporaneous average wait exceeds a pre-specified threshold
- `days34_indicator`: 1 if experiment day is 3 or 4

J. Practical note

The cleanest workflow is to extract the raw platform and delayed-test logs first, then collapse them to the student level using the variable definitions above. The final analysis spreadsheet should be a student-level file with one row per student and the variables needed for the main tables and regressions.