Navvy+

Impact Evaluation of Pearson Navvy Math in Arizona School Districts

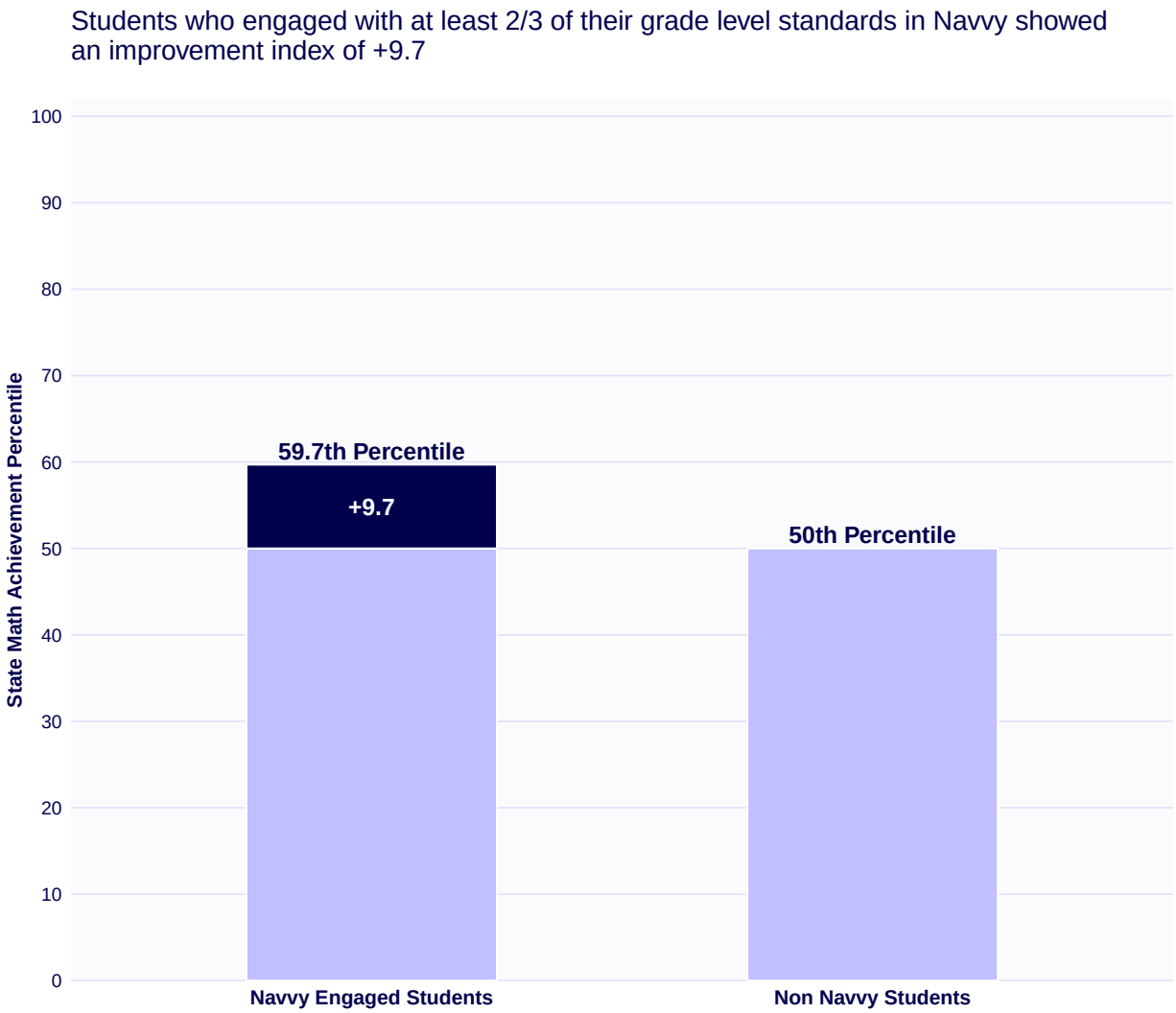
2024–2025 Academic Year

Laine Bradshaw, PhD
Athul Sudheesh, PhD
Madeline Schellman, PhD
Ruchi Sachdeva, PhD

Executive Summary

A study of Navvy's initial adoption among 4th to 8th-grade students in Arizona during the 2024–2025 academic year found a statistically significant positive impact on math proficiency. Students who engaged with at least 2/3 of the available grade-level standards in Navvy demonstrated meaningfully higher growth on the Arizona's Academic Standards Assessment (AASA) compared to their peers in a matched control group. The study also found that higher levels of Navvy Engagement were associated with greater learning gains. In addition, among students who were Not Proficient in the spring 2024, those with higher levels of Navvy Engagement were more likely to reach Proficient status in spring 2025 than matched peers who did not use Navvy.

Together, these findings provide strong evidence from a quasi-experimental design meeting What Works Clearinghouse standards that Navvy is an effective tool for supporting mathematics learning.



Introduction

Navvy is a K-12 formative classroom assessment solution within the Pearson Assessment for Learning Suite (PALS) that encourages instructor-assisted personalized learning and nurtures healthy learning mindsets by providing rigorous and engaging standards-aligned tasks that produce granular, reliable, and timely learning evidence. Developed in collaboration with measurement scientists, assessment designers, educators, and educational leaders, Navvy leverages learning and measurement science to deliver brief, formative assessments with the rigor of large-scale assessments. By providing educators with granular, psychometrically defensible data, the Navvy system highlights specific competency learning and gaps at the standard level. This actionable data powers the celebration of specific learning and enables real-time instructional adjustments, allowing teachers to close learning gaps quickly and accelerate overall student growth.

The Navvy system integrates two primary components to support instructor-assisted personalized learning and nurture healthy learning mindsets: formative assessments, known as *Checks*, and a repository of instructional materials called the *Learning Library*. The Checks are delivered in two formats. *Competency Checks* are brief, secure assessments, typically composed of six to eight items that are intentionally designed to reflect the breadth of content and depth of cognitive complexity the standard requires, enabling reliable diagnosis of student competency on a single academic standard. Second, *Practice Checks* are non-secure, variable-length assessments that provide students with standards-aligned practice tailored to their individual learning needs. These can be created by educators or generated automatically by Navvy's intelligent recommendation layer.

Complementing these assessments, the *Learning Library* contains high-quality instructional materials (HQIM) designed to support the learning of individual standards, helping to supplement instruction and facilitate personalized learning or targeted instruction. The Navvy system provides Checks across K-12 grade-levels in English Language Arts (ELA), Math, Science, and Social Studies, with the Learning Library resources currently available for ELA and Math.

Navvy delivers assessment results through dynamic reporting dashboards. To provide insight needed to personalize learning, teachers and administrators receive student- and classroom-level insights that identify which standards, and which parts of standards, students have learned and which they need more support to learn (see, for example, Figure 1). Additional insights for standards-level learning aggregated at the school-, district-, and state- levels are provided for administrators (see Figure 2). Students access a learner-facing progress dashboard that frames learning as a mission to earn micro-credentials (badges) for each standard and provides immediate access to results and feedback to support meta-cognitive reflection and assessment-as-learning practices (see Figure 3).

Initial Adoption of Navvy in Arizona schools

Navvy was introduced in Arizona in the fall of 2024, with the state providing school districts the opportunity to opt in to use Navvy as a resource. Participating schools began using Navvy in late 2024 and early 2025. Because participation was optional, engagement varied across schools and grade-levels during this initial implementation period. By the end of the 2024–2025 school year, Navvy had expanded to 82 schools across 60 districts, serving over 18,000 students and 600 teachers.

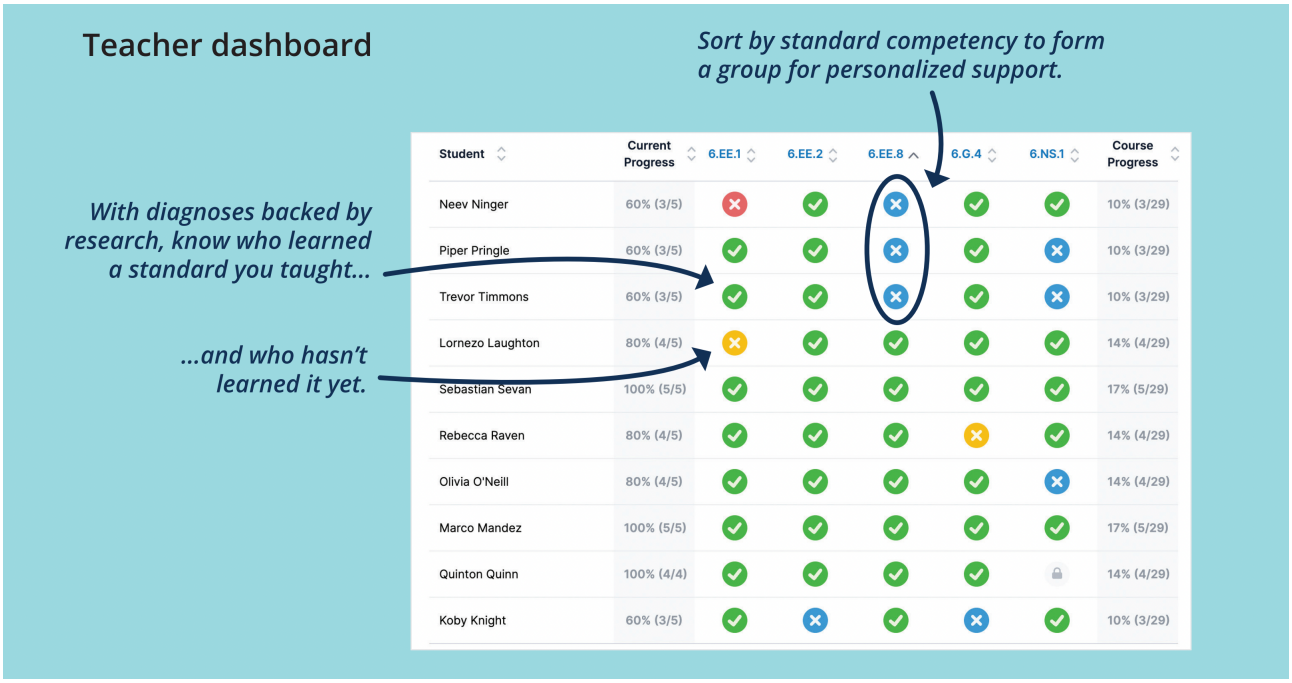


Figure 1: Snippet of Navvy dashboard for a teacher, showing student learning profiles that detail competency by academic content standard for a math class.

Goals of the Study

The purpose of this study was to evaluate the impact of Navvy on student learning outcomes within this initial cohort of adopting schools in Arizona. Specifically, the following research questions (RQs) guided the design and analysis of the study.

RQ 1 What is the effect of students engaging with Navvy on at least 2/3 of grade-level standards available in Navvy (i.e., Navvy Engagement) on mathematics achievement among students in Arizona, as measured by the Arizona statewide assessment (Arizona Academic Standards Assessment, AASA) scaled scores in the 2024–2025 school year?

RQ2 Among Navvy Engaged students, is there what is known as a 'dose-response' relationship, where higher levels of Navvy Engagement correspond to greater learning gains?

Methods

Research Design

This study used a student-level quasi-experimental design to estimate the effect of Navvy Engagement on students' mathematics achievement during Navvy's initial implementation in Arizona. Participation was opt-in, and implementation varied across schools, teachers, and grade-levels, so the analysis focused on the student level rather than the school level. Because students were not randomly assigned to treatment and control groups, we used propensity score matching to reduce pre-existing differences between groups and establish baseline equivalence before estimating treatment effects. The analytic model also accounted for differences across schools and grades to better isolate the effect of Navvy Engagement on students' math

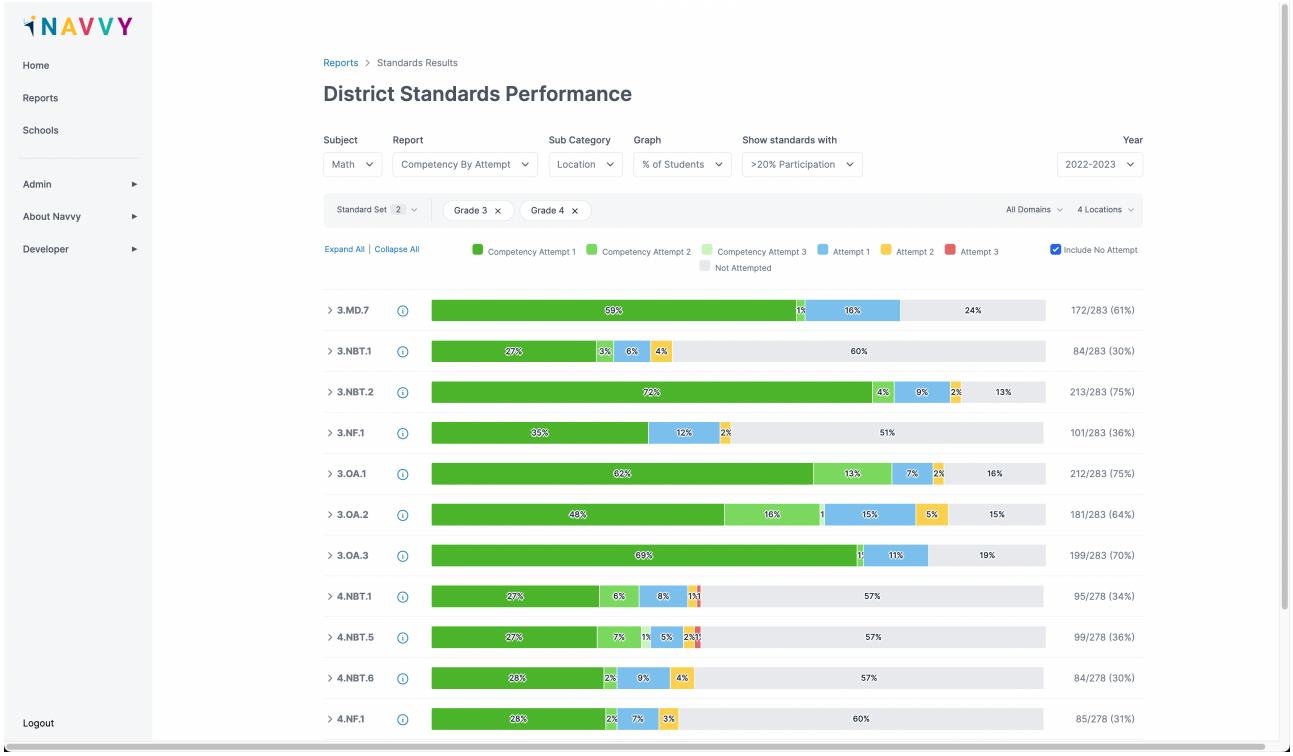


Figure 2: Snippet of a sample administrator level dashboard in Navvy. This view aggregates student competency data from multiple schools to provide insights into district-wide performance and inform strategic decision-making.

achievement.

Analytical Sample

Students were included in the analytic sample if they met all of the following criteria: (a) they were enrolled in Grades 4–8; (b) they could be matched across Navvy and AASA data sources; (c) they had AASA scaled scores for both the prior year (2023–2024) and current year (2024–2025); (d) they had complete demographic information required for the analysis; and (e) they attended a school with at least 50 eligible students to support stable estimation.

Navvy response data from the implementation year included 14,439 students across 84 schools in Grades 4–8. After merging with AASA records and applying the requirement for two years of AASA scores, the sample yielded 10,510 students from 60 schools. Applying the remaining inclusion criteria yielded a final analytic sample of 6,141 students from 47 schools. These students were eligible to be assigned to either the Navvy Engagement treatment group or the control group.

Navvy Engagement Treatment and Comparison Group Definition

Students were assigned to the Navvy Engagement treatment group if they meaningfully engaged with either a Competency Check or a Practice Check for at least two-thirds of the grade-level standards available in Navvy (see Table 1). Meaningful engagement was defined as submitting a Check as an effortful responder, such that the time spent on the Check was not flagged

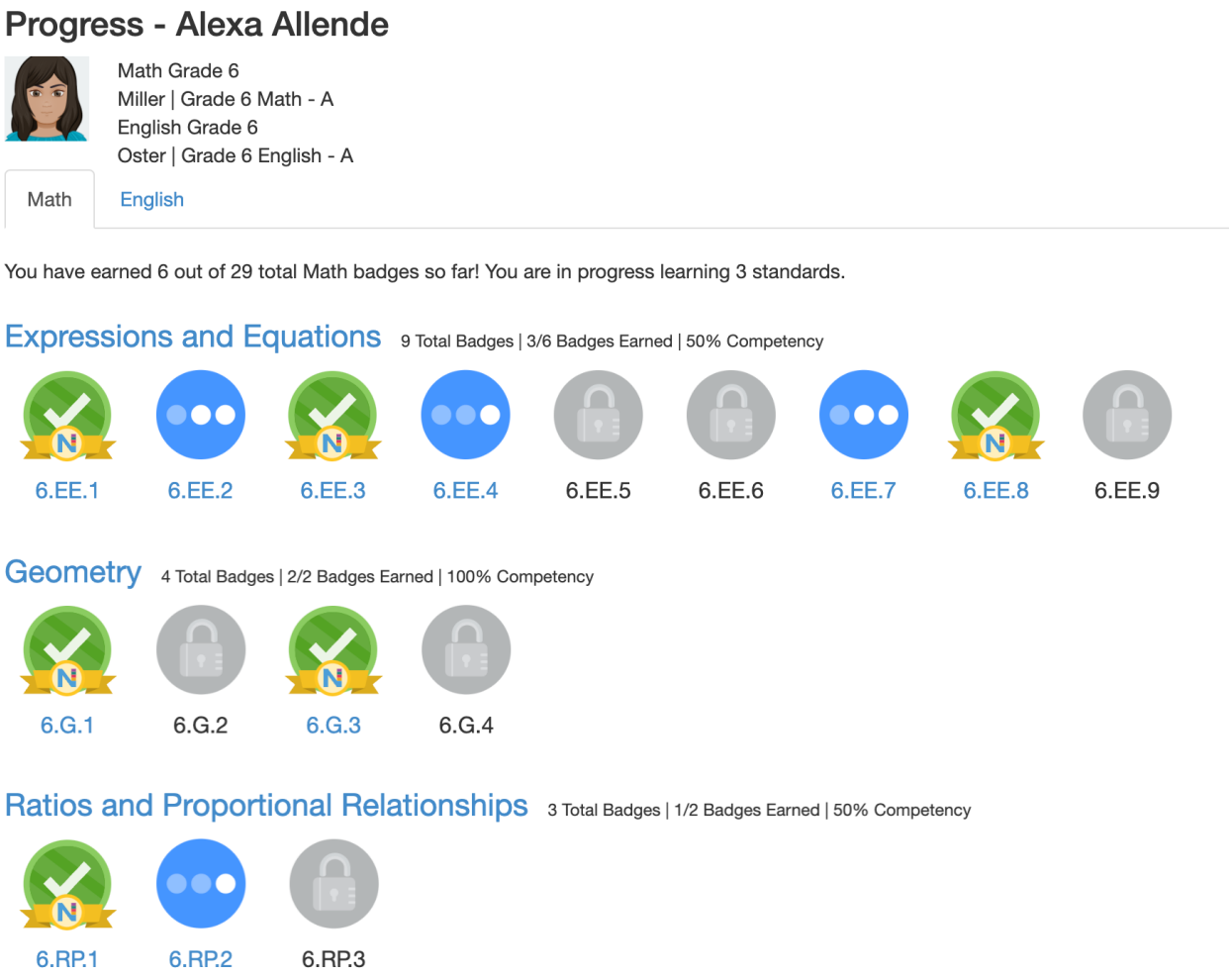


Figure 3: Snippet of student-facing dashboard providing badge-based progress toward micro-credentials and immediate results to support assessment-as-learning.

for rapid responding. Students with no recorded Navvy assessment activity were eligible for the control group. Applying these definitions yielded the focal comparison sample used in the matched analysis.

Table 1: Total number of grade-level mathematics standards with Navvy assessments available.

Grade Level	Math Standards
3	26
4	29
5	26
6	29
7	23
8	29

Matching and Baseline Equivalence

Propensity score matching (Rubin, 1997) was used with a nearest neighbor algorithm to form a control group that was statistically similar to the treatment group on prior scaled scores and key demographic variables. The `MatchIt` R package (Ho et al., 2011) was used to prepare the propensity score-matched analytic sample.

Post-matching balance was evaluated using standardized mean differences (SMDs). According to What Works Clearinghouse (2022) criteria, baseline equivalence is established when the SMD is below 0.05. The matched sample had an SMD of 0.018, indicating that the groups were well matched. Figure 4 summarizes covariate balance before and after matching, and Tables 2 and 3 present the corresponding descriptive statistics and SMDs before and after matching, respectively.

Table 2: Demographic characteristics and standardized mean differences before matching

	Treatment			Control				
Variables	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Mean Difference**	**SMD**
Prior Math (SY23–24)	193	3596.79	59.80	261536	3576.15	59.42	20.64	0.347
Hispanic %	110	56.99	49.64	125499	47.99	49.96	9.01	0.180
American Indian %	6	3.11	17.40	16935	6.48	24.61	-3.37	-0.137
Asian %	6	3.11	17.40	15178	5.80	23.38	-2.69	-0.115
African American %	20	10.36	30.56	24345	9.31	29.06	1.05	0.036
White %	165	85.49	35.31	216570	82.81	37.73	2.69	0.071
Native Hawaiian %	5	2.59	15.93	3234	1.24	11.05	1.35	0.122
English Language Learner %	14	7.25	26.01	25093	9.59	29.45	-2.34	-0.079
Special Education %	25	12.95	33.67	36851	14.09	34.79	-1.14	-0.033
Female %	88	45.60	49.94	128370	49.08	49.99	-3.49	-0.070

Table 3: Demographic characteristics and standardized mean differences after matching

	Treatment			Control				
Variables	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Mean Difference**	**SMD**
Prior Math (SY23–24)	193	3596.79	59.80	193	3597.90	60.09	-1.11	-0.019
Hispanic %	110	56.99	49.64	109	56.48	49.71	0.52	0.010
American Indian %	6	3.11	17.40	5	2.59	15.93	0.52	0.031
Asian %	6	3.11	17.40	7	3.63	18.74	-0.52	-0.029
African American %	20	10.36	30.56	20	10.36	30.56	0.00	0.000
White %	165	85.49	35.31	166	86.01	34.78	-0.52	-0.015
Native Hawaiian %	5	2.59	15.93	5	2.59	15.93	0.00	0.000
English Language Learner %	14	7.25	26.01	13	6.74	25.13	0.52	0.020
Special Education %	25	12.95	33.67	25	12.95	33.67	0.00	0.000
Female %	88	45.60	49.94	88	45.60	49.94	0.00	0.000

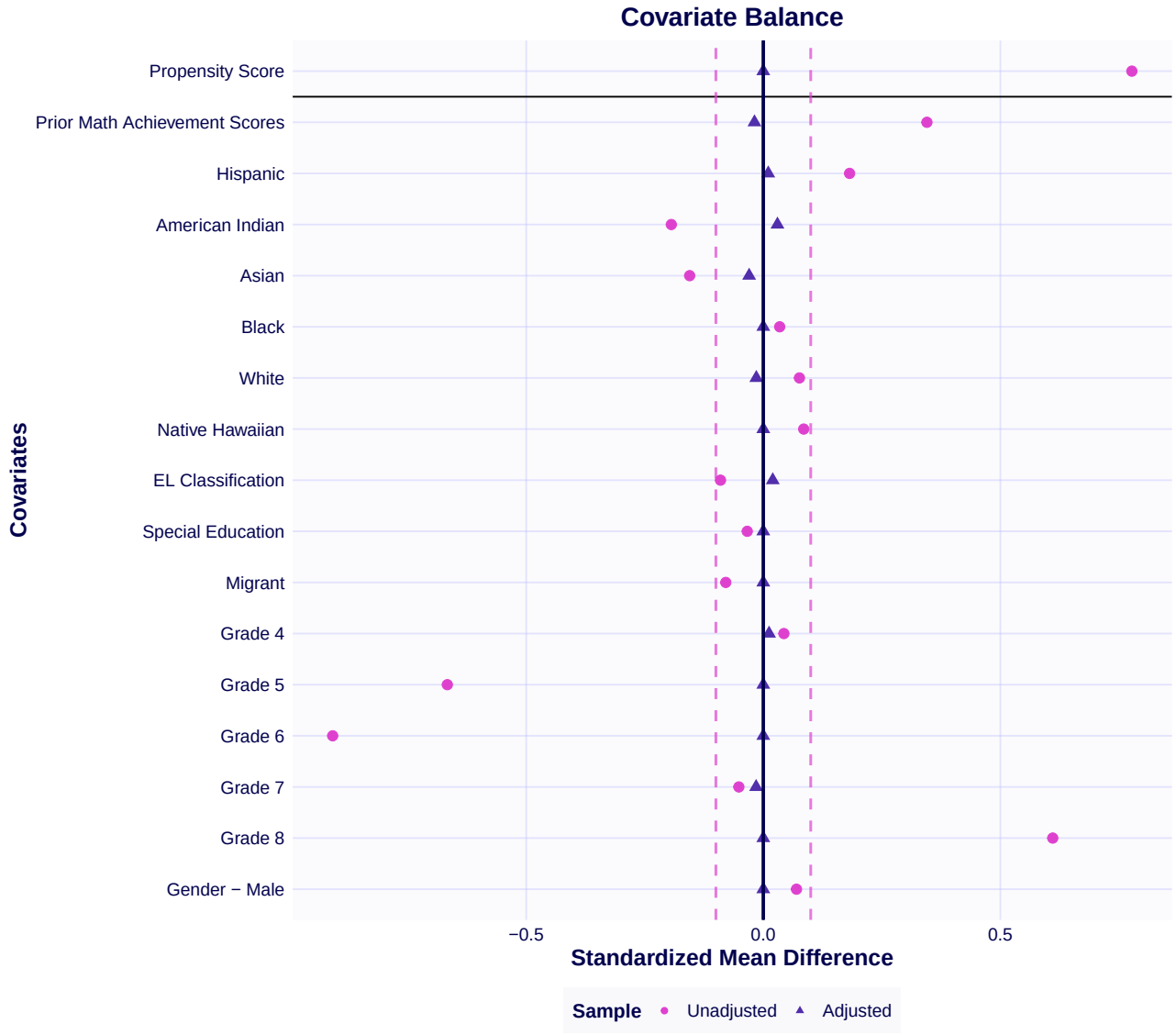


Figure 4: Covariate balance before (unadjusted) and after (adjusted) matching.

Analytical Model

The effect of Navvy Engagement on mathematics achievement was estimated using a two-level hierarchical linear model (Bryk & Raudenbush, 1987) with random effects for schools and grades. The model specification is:

$$Y_{ijg} = \beta_0 + \beta_1 X_{ijg} + \beta_2 T_{ijg} + u_j + v_g + \epsilon_{ijg} \tag{1}$$

where

- Y_{ijg} = State mathematics achievement score for student i in school j, grade g (Spring 2025)
- X_{ijg} = Prior year state mathematics achievement score for student i (Spring 2024)
- T_{ijg} = Treatment indicator: $T = 1$ if student has engaged with $\geq 2/3$ Navvy standards, $T = 0$ otherwise
- β_0 = Intercept

- β_1 = Coefficient for prior achievement
- β_2 = Treatment effect of Navvy Engagement on student achievement
- u_j = School-specific random effect (accounts for unobserved school characteristics)
- v_g = Grade-specific random effect (accounts for grade-level differences)
- ϵ_{ijg} = Student-level error term

The model parameters were estimated using the `lme4` R package (Bates et al., 2015). We included a random effect for grade-level because Grades 4–8 reflect distinct developmental and curricular stages that are associated with systematic differences in baseline performance (see Figure5). As illustrated in Figure 5, the distribution of prior year mathematics achievement scores shifts upward with grade-level (including higher grade-level means), consistent with clustering by grade. This pattern supports modeling grade as a random effect to capture between-grade heterogeneity in achievement

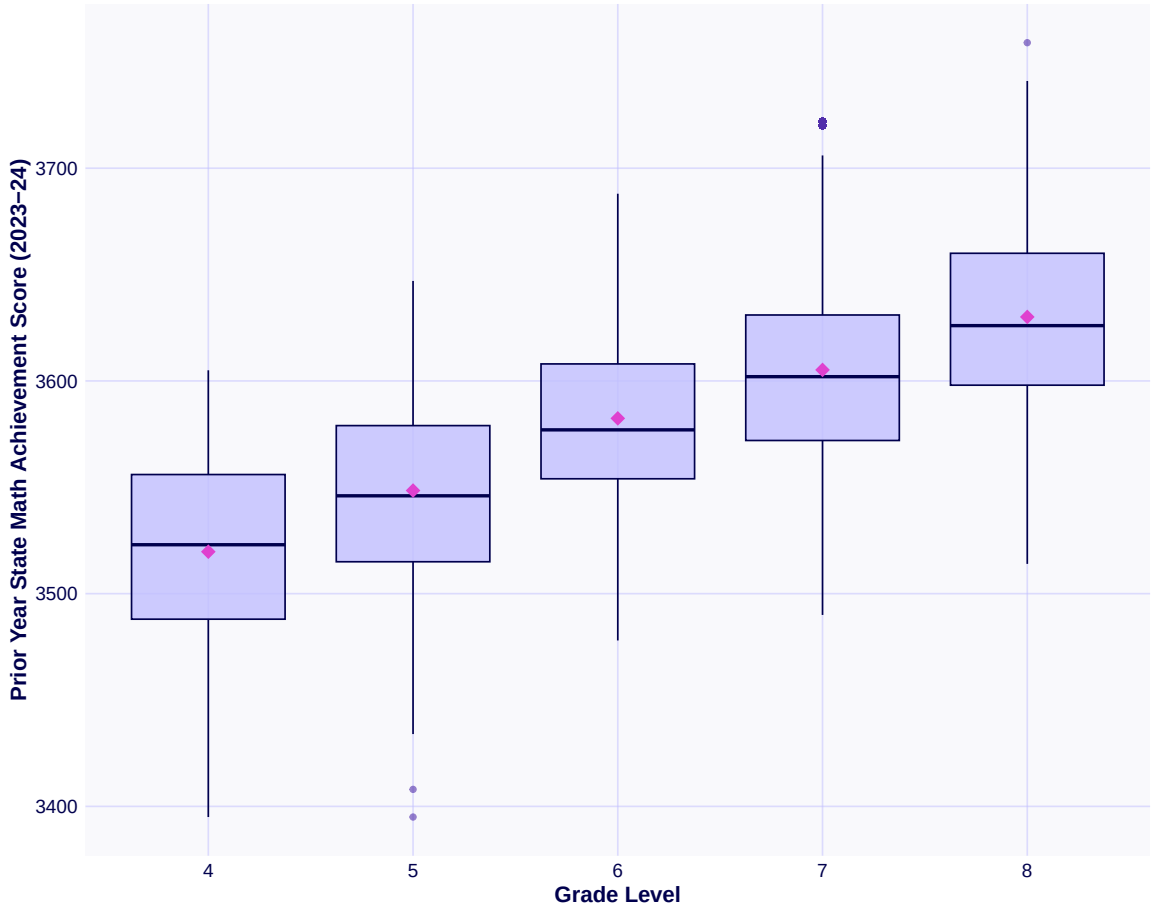


Figure 5: Prior Year Math Achievement Score Distribution by Grade Level. Pink diamonds indicate grade-level means.

To empirically validate the necessity of this model structure, several diagnostic procedures were conducted. The extent of outcome clustering within grade-levels was quantified by calculating the intraclass correlation coefficient (ICC). The proportion of total outcome variance attributable to grade-level differences was further examined by means of a variance decomposition analysis. Finally, to formally justify the inclusion of grade-level random effects, a likelihood-ratio test was conducted.

Analysis Plan

To address **RQ1**, the treatment group was defined as students who completed a Navvy Check for at least 2/3 of the unique grade-level standards available in the Navvy system. The effects of Navvy Engagement were estimated using the analytical model described in Equation 1.

To address **RQ 2**, we conducted a dose–response analysis by re-estimating Equation 1 across 17 analytic samples, systematically varying the engagement threshold from 5% to 85% of grade-level standards in 5-percentage-point increments. In addition, we examined achievement growth by comparing the proportion of students who transitioned from Not Proficient on the prior year's AASA (2023–2024) to Proficient on the current year's AASA (2024–2025) in matched treatment and control groups at each engagement threshold.

The effect sizes from the analysis were converted to Hedge's *g* and to the improvement index for better interpretability as recommended by the What Works Clearinghouse (2022).

Results

Table 4: Two-Level Mixed-Effects Model Predicting Student Achievement.

	Current Year Math Achievement Score			
Predictor	*β*	*SE*	*95% CI*	*p*
Intercept	446.63	99.59	250.81 – 642.44	**<0.001**
Prior Year Math Achievement Score	0.88	0.03	0.83 – 0.94	**<0.001**
Treatment (Navvy Engagement)	14.84	4.70	5.59 – 24.09	**0.002**
Random Effects				
σ^2	434.60			
$\tau_{00\ Schools}$	111.47			
$\tau_{00\ Grade}$	21.84			
ICC	0.23			
$N_{Schools}$	92			
N_{Grade}	5			
Observations	386			
Marginal R^2 / Conditional R^2	0.833 / 0.872			

Note. The model includes treatment status and prior year math achievement score as fixed effects, with random intercepts for schools and grades.

Table 4 shows the parameter estimates from the two-level hierarchical mixed-effects model used to evaluate the impact of Navvy Engagement on student achievement. The model included fixed effects for treatment status and prior year math achievement score, with random inter-

cepts for schools and grades to account for data clustering. The analysis was conducted on a sample of 386 observations across 92 schools and 5 grades.

Results: Navvy Engagement Leads to Statistically Significant Gains

Results showed that Navvy Engagement had a statistically significant positive effect on student mathematics achievement ($\beta = 14.84$, 95% CI [5.59, 24.09], $p < 0.01$). In practical terms, **students who engaged with Navvy on at least two-thirds of the available grade-level standards statistically significantly outperformed a matched group of students not using Navvy**, after accounting for prior year state achievement and clustering by school and grade. The treatment estimate translates to an effect size of Hedge's $g = 0.245$ and to an improvement index of +9.7. This finding is notable in educational contexts because effects on broad, independent standardized achievement outcomes are typically smaller than effects observed on more proximal or researcher-designed measures, making an effect of this magnitude meaningful under real-world implementation conditions (Kraft, 2020).

Results indicated the use of a mixed-effects model was justified by the significant amount of variance attributable to the grouping structures. The intraclass correlation coefficient (ICC) was 0.23, calculated from the variance components for school ($\tau_{00} = 111.47$), grade ($\tau_{00} = 21.84$), and the residual variance ($\sigma^2 = 434.60$). Decomposing this reveals that school-level differences accounted for approximately 19.6% of the unexplained variance, while grade-level differences accounted for 3.8%. This combined ICC value signifies that roughly 23% of the total variance in student scores is attributable to these between-school and between-grade contextual factors. A likelihood ratio test further confirmed that the model including a random effect for grade provided a significantly better fit than a model without it ($\chi^2(1)$=18.78, $p < .001$). Additionally, the Akaike Information Criterion (AIC) was lower for the model with grade-level random effects (3497.7) compared to the reduced model (3512.5), further confirming that accounting for the nested structure of the grade provides a more accurate estimation of treatment effects.

The marginal R^2, which represents the variance explained by the fixed effects – specifically, the prior year achievement score and Navvy treatment – was 0.833. When also accounting for the grouping structure, such as the nesting of students within schools and grades, the conditional R^2 rose to 0.872. Together, these student-level predictors and school-level and grade-level differences account for 87.2% of the variance in achievement scores, demonstrating that the model had substantial explanatory power.

Results: Higher Engagement Yields Higher Learning Gains

The results from the dose-response analysis revealed that there was a strong association between Navvy Engagement and learning gains: **as students engaged with higher percentage of the Navvy standards, the positive impact on their math achievement scores also increased** (see Figure 6). The treatment effect is statistically significant ($p \leq 0.05$) across nearly all thresholds analyzed (from 5% up to 80%). This indicates that Navvy Engagement is beneficial even at lower levels of exposure, but the benefits compound with increased exposure.

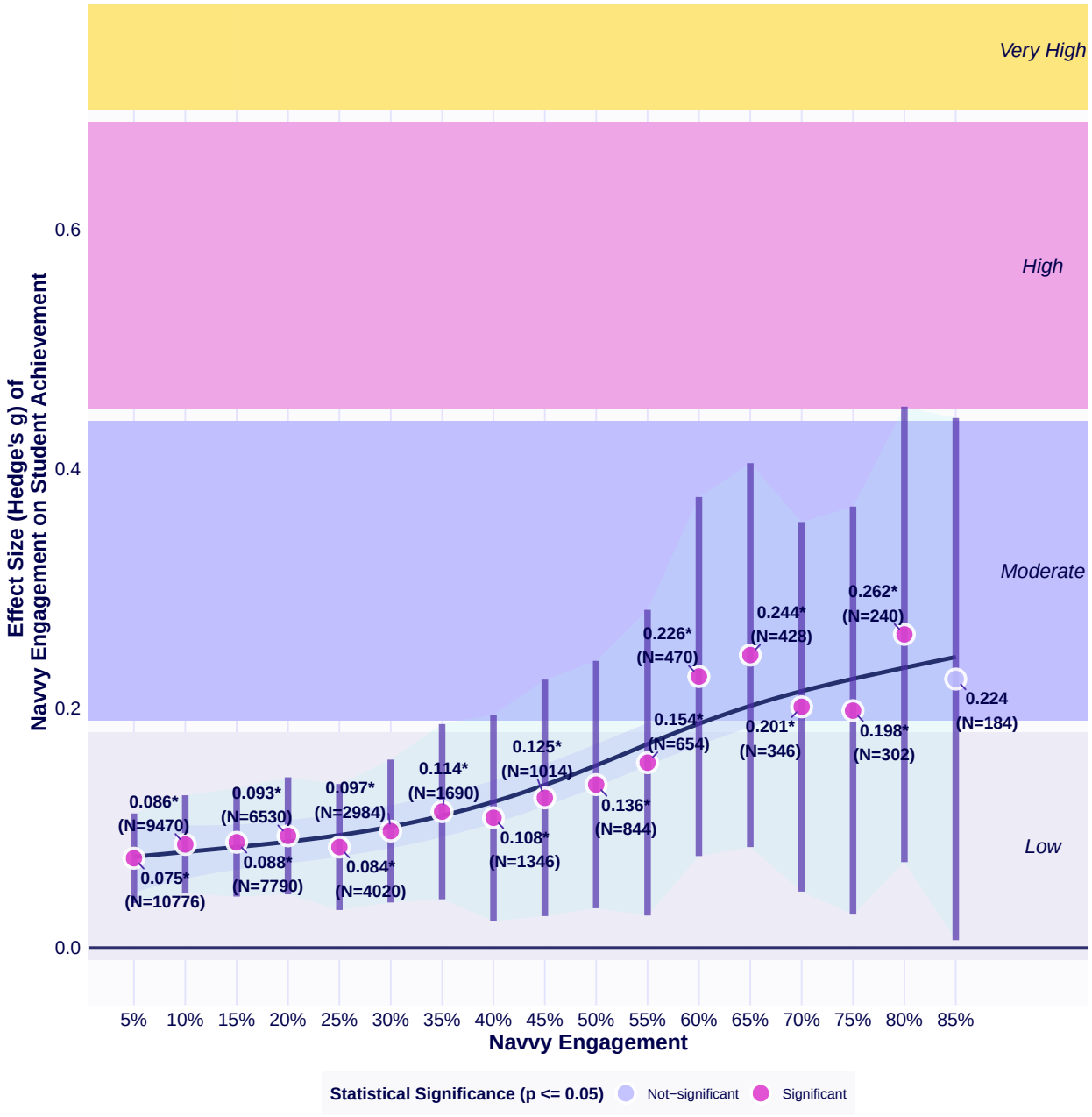


Figure 6: Dosage-Response Relationship Between Navvy Engagement and Learning Gains.

Results: Navvy Helps Close Achievement Gaps

A key milestone in state assessment for students is reaching the Proficient level set by their state. To evaluate the practical significance of the impact of Navvy Engagement on student achievement, we calculated the transition rate of students in our matched sample moving from 'Not Proficient' (AASA performance levels 1 and 2) status in the prior year state achievement to 'Proficient' (AASA performance levels 3 and 4) status in the current year across varying thresholds of engagement intensity (see Figure 7). The results demonstrate that **higher engagement with Navvy substantially increases the likelihood of a student moving from Not-proficient to Proficient status.**

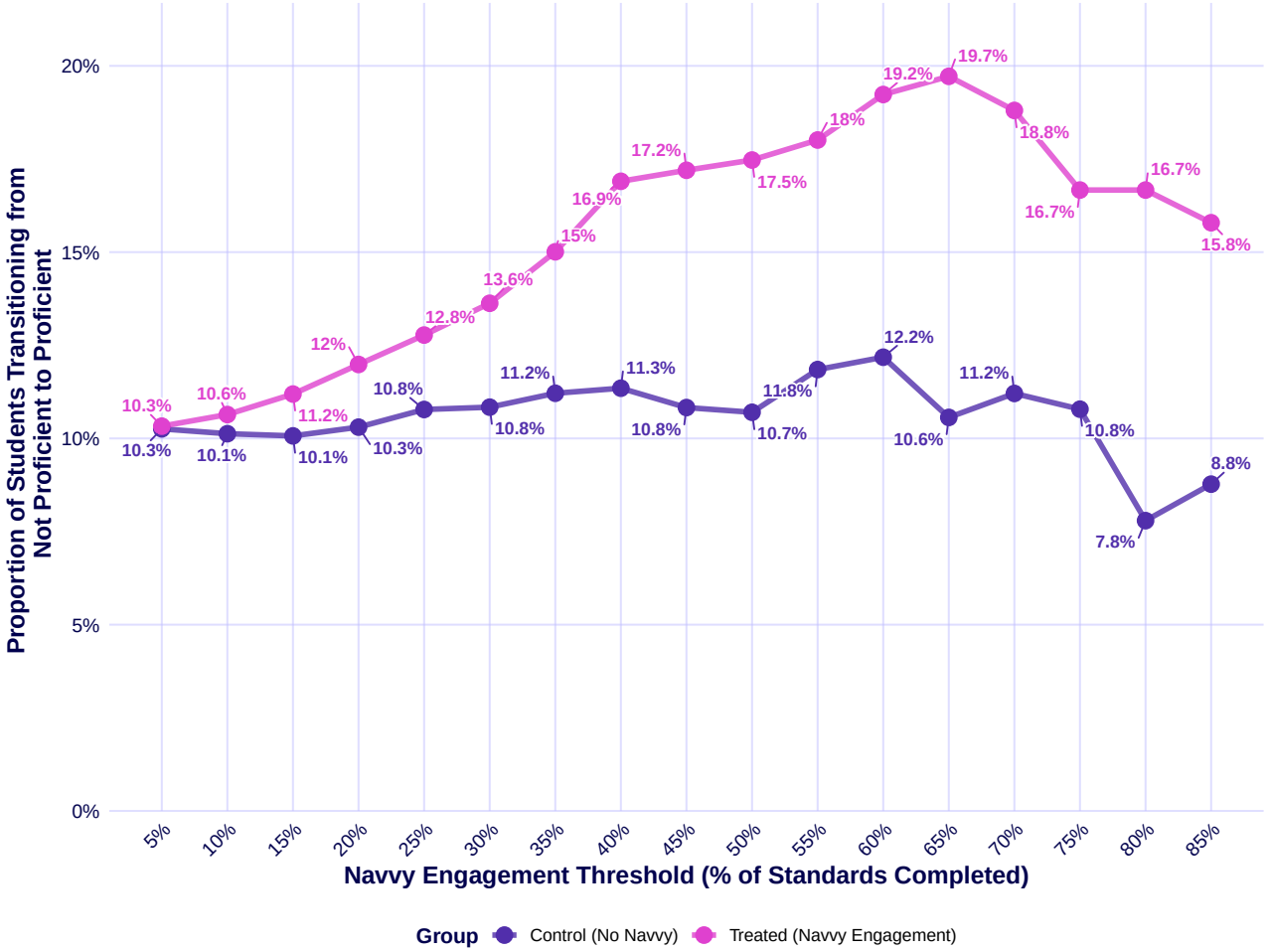


Figure 7: Proportion of students moving from "Not Proficient" in spring 2024 to "Proficient" in the spring 2025.

Across engagement thresholds, **the gap between students with Navvy Engagement (treatment group) and the matched control group of students not using Navvy widened as students completed more Navvy standards**, indicating that higher engagement was associated with a greater probability of closing achievement gaps. At low engagement levels (5–15%), the difference between the Navvy Engagement (treatment) and matched group of students not using Navvy (control group) is negligible, with both groups showing a proficiency transition rate of approximately 10%. However, as engagement increases, a clear divergence emerges. By the 35% engagement threshold, the gap widens significantly: 15.0% of students with Navvy Engagement
achieved proficiency compared to only 11.2% of the matched control group of students who did not use Navvy.

The strongest benefit appeared between the 60% and 70% engagement thresholds, where nearly 1 in 5 (19.7%) Navvy Engaged students who were Not Proficient in spring 2024 reached Proficient in spring 2025, compared with approximately 1 in 10 (11.2%) matched control students not using Navvy. At the peak effectiveness range (65% threshold), students engaging with Navvy were nearly twice as likely (19.7% vs. 10.6%) to achieve proficiency as their matched peers who did not use the Navvy system.

Future Research

This evaluation used a rigorous quasi-experimental design to answer two important questions: whether Navvy Engagement was associated with improved mathematics achievement, and whether higher levels of engagement were associated with stronger learning outcomes. The findings provide strong evidence that Navvy can support student learning in mathematics under real-world implementation conditions. Strong evidence is important to ensure learning systems are doing what they are intended to do: Help students learn. A single study doesn't tell the whole story of a learning system. Future work can answer additional questions. This section discusses how additional questions can be asked and answered.

This study used a quasi-experimental design, which is well suited to evaluating educational programs in authentic school settings where random assignment is often difficult or not feasible. In this design, students who engaged with Navvy were compared with a matched group of students not using Navvy who were similar on prior mathematics achievement and observed demographic characteristics. This approach strengthens causal interpretation, although, as in all observational studies, future work could further reduce uncertainty by incorporating additional information about student, teacher, and school contexts. Randomized controlled trials can provide an even stronger basis for causal inference, but they are often challenging to implement in applied school settings because they require agreement on assignment procedures, stable implementation conditions, and sufficient operational flexibility across participating sites.

Future studies could extend the present work in several ways. First, they could incorporate additional contextual variables that were not available here, such as teacher implementation practices, use of other instructional supports, attendance patterns, or school-level policies. If characteristics such as these are related to both Navvy Engagement and achievement outcomes, including them would allow even more precise estimation of Navvy's effects.

Second, future work could examine implementation more directly. The current study measured Navvy Engagement as the proportion of available grade-level standards with which students meaningfully engaged. Future research can gather richer implementation variables, such as how teachers engaged with Navvy, whether Learning Library resources were used alongside Checks, and how engagement was distributed across the school year. These data would help clarify which implementation patterns are most strongly associated with improved outcomes.

Third, future studies could broaden the range of outcomes examined. In this evaluation, student achievement was measured by the end-of-year state mathematics assessment, which is an important and independent outcome. Future research could also investigate whether Navvy influences other outcomes that matter for learning, such as student confidence, persistence, engagement with challenging materials as well as teacher decision-making during instruction.

Discussion

This study provides strong evidence that Navvy Engagement was associated with improved student mathematics achievement. The What Works Clearinghouse (WWC) identifies several key features of rigorous quasi-experimental research, including baseline equivalence between treatment and comparison groups, use of an independent outcome measure, and appropriate analytic methods to account for clustering in the data. This study met each of those expectations.

Baseline equivalence was established through propensity score matching, the outcome was measured using an independent state standardized assessment, and hierarchical linear modeling was used to account for school-level clustering. Together, these design features support the internal validity of the findings and position this study as meeting the criteria for ESSA Tier 2 (Moderate Evidence). **These findings provide validated, empirical support that implementation of Pearson Navvy is associated with meaningful improvements in student mathematics proficiency, consistent with its design as an integrated system supporting both assessment-for and assessment-as-learning.**

Why Navvy Works?

Navvy is designed as a measurement-informed learning system that integrates psychometric rigor, learning science, and instructional practice. Rather than centering on single scores or static evaluations, the system is built to generate precise, actionable evidence that supports educator decision-making and student learning within the instructional cycle. Navvy's theory of action centers on an iterative learning cycle in which high-quality tasks both generate evidence of learning and support its development. Each assessment engagement provides an opportunity to strengthen understanding while producing feedback to inform reflection and next instructional steps.

Core Design Principles behind Navvy System

1. **Competency Diagnosis over Composite Scores**

   Navvy prioritizes competency diagnosis at the level of individual state standards rather than relying on composite scores. Checks are constructed to reflect each standard's components and intended cognitive complexity (depth-of-knowledge; DOK), yielding granular evidence at the student, class, school, and district levels. This diagnostic approach enables educators to identify not only whether a standard has been learned, but where within the standard additional support is needed.

2. **Assessment *for* Learning over Assessment *of* Learning**

   Consistent with formative assessment frameworks, Navvy is designed to inform and advance learning rather than sort students based on overall scores. Students may engage in multiple attempts on Competency Checks, with the expectation that additional instruction occurs between attempts. Teachers may also enable answer review on Practice Checks. Student dashboards present current and prior results, including component- and DOK-level indicators, to support goal setting and self-monitoring. Feedback structures, including progress-focused encouragement notes, are designed to attribute growth to strategy and effort, consistent with research on feedback and learning mindsets (Dweck, 2006).

3. **Autonomy over Automation**

   Navvy does not prescribe a fixed instructional pathway. Educators retain flexibility in how the system is integrated into classroom practice, including when to assess, which standards to combine, and how to respond to the resulting evidence. Short, pre-built assessments (typically 6–8 items) reduce authoring overhead while preserving instructional agency. Retake opportunities are embedded by design to align assessment with student readiness to demonstrate competency.

4. **Assessment as Learning**

   Navvy's assessment tasks are intentionally designed not only to measure learning, but also to support deeper learning and long-term retention. Each task is aligned to grade-level standards and constructed to elicit the level of thinking those standards require, including the appropriate depth of knowledge (DOK). As students work through these tasks, they retrieve prior knowledge, apply concepts, and make connections across ideas—processes that learning science identifies as central to durable learning. By embedding principles such as retrieval practice and interleaving into assessment experiences, Navvy positions assessment engagement itself as a meaningful part of the learning process.

Taken together, these principles provide a coherent account of why effects appear when Navvy is implemented with sufficient breadth and regularity. These effects emerge through the integration of diagnostic precision, formative use, educator agency, and assessment-as-learning design.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software*, *67*, 1–48.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological bulletin*, *101*(1), 147.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random house.

Ho, D., Imai, K., King, G., & Stuart, E. A. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of statistical software*, *42*, 1–28.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational researcher*, *49*(4), 241–253.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, *127*, 757–763.

What Works Clearinghouse. (2022). *What works clearinghouse: Standards handbook (version 5.0)* (tech. rep.). Institute of Education Sciences, U.S. Department of Education. http://whatworks.ed.gov