The Effect of Navvy on Arizona Student Mathematics Achievement

Authors: Hye-Jeong Choi
Caroline Wiley
Yachen Luo

Date: December 16, 2025

Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314 | Phone: 703.549.3611 | humrro.org

The Effect of Navvy on Student Mathematics Achievement (Arizona)

Table of Contents

List of Tables

List of Figures

Navvy Impact on Student Mathematics Achievement (Arizona)

Introduction

The Human Resources Research Organization (HumRRO) contracted with Pearson to independently evaluate Navvy impact on student achievement in mathematics in grades 4 through 8. Navvy is a suite of standards-based classroom assessments designed to provide real-time, actionable, and accurate results to inform personalized instruction in English Language Arts (ELA) and mathematics. Navvy includes Competency Checks, which are secure, psychometrically sound competency assessments, and Practice Checks, which are non-secure, flexible items. The purpose of the study is to determine to what extent students who participate in any Navvy (Competency Checks or Practice Checks) outperform students who do not participate in any Navvy on the Arizona state summative assessment in mathematics.

Research questions

The purpose of this study was to evaluate the effect of Navvy on students' achievement in mathematics for grades 4 through 8. The primary research questions were:

- RQ1. What are the effects of Navvy usage on mathematics achievement among students in grades 4−8 in Arizona, as measured by Arizona's Academic Standards Assessment (AASA) in the 2024−25 school year?
- RQ2. Among students who use Navvy, what is the relationship between the level of participation and their mathematics achievement?

Methods

Research design

This study employed a Quasi-Experimental Design (QED) to evaluate the effects of Navvy in grades 4-8 on the state summative mathematics assessment (Arizona's Academic Standards Assessment [AASA]), following appropriate methods to meet What Works Clearinghouse (WWC) standards with reservations.

In the state of Arizona, Local Education Agencies (LEAs) have full autonomy to implement Navvy according to their specific needs. Implementation can range from district- or school-wide adoption to targeted deployment for specific student groups. These diverse implementation approaches align with Pearson's intended flexibility. As such, we treated Navvy as a student-level intervention, and our study employed the individual student as the unit of analysis.

Data

Following initial data cleaning and processing by Pearson, the prepared datasets were transmitted to HumRRO for analysis. The final data sources included Navvy usage data, state assessment scale scores (SY2023-2024 and SY2024-2025), and student demographic information. Pearson provided a data set with 267,652 grade 4 through 8 student records from 1,287 schools. Students were classified into the treatment group if they completed Navvy assessments for over 65% of standards, while those who never used Navvy formed the comparison group. Students who did not belong to either the comparison or treatment group

were excluded from this analysis. The initial sample consisted of 214 treatment and 261,536 comparison students.

To ensure a fair representation of the treatment effect, we used *grade-mean* centering, in which students' prior-year scores were centered around the mean score of their grade. This approach removes all the between-grade differences in prior achievement, ensuring that each student is compared only to peers in the same grade. Because the AASA is vertically scaled, the treatment effect can therefore be interpreted as an average effect within grade, after accounting for school-level differences.

Matching

What Works Clearinghouse (WWC) recognizes statistical matching as an acceptable baseline adjustment strategy (WWC, 2022) to support causal inference. Using the R package MatchIt (Ho et al., 2018), we employed nearest neighbor matching with a 1:1 matching ratio. We balanced treatment and comparison students by matching students' SY2023-24 mathematics scale scores and demographic information (grade level, race/ethnicity, English Language Learner [ELL] status, Student with Disabilities [SWD] status, migrant status, and gender).

We used student-level covariates to match comparison students. The covariates included were:

- **Prior Score**: SY 2023-24 mathematics scale score
- **Female%**: Percent of students identifying as Female
- **Hispanic%**: Percent of students identifying as Hispanic
- **American Indian%**: Percent of students identifying as American Indian
- **Asian%**: Percent of students identifying as Asian
- **African American%**: Percent of students identifying as African American
- **White%**: Percent of students identifying as White
- **Native Hawaiian%**: Percent of students identifying as Native Hawaiian
- **ELL%**: Percent of English language learners
- **SWD%**: Percent of students with disabilities
- **Migrant%**: Percent of students identifying as a part of a migrant education program

Before matching, substantial imbalances were evident between the treatment and comparison schools in the distribution of prior scores and student demographic characteristics. Table 1 shows the baseline equivalence prior to matching.[1]

The matching procedure achieved satisfactory covariate balance across covariates with effect sizes below 0.05, suggesting no further statistical adjustment is required (WWC, 2022)[2]. Table 2 shows the baseline equivalence after matching.

[1] Although the matching procedure used a grade-centered version of the prior scores, all descriptive statistics (means and Standard Deviations [SDs]) for the prior scores are reported on the original raw scale.

[2] What Works Clearinghouse Handbook version 5 (2022) recommends using the pooled raw student-level standard deviation to calculate the effect size. That is, WWC standardizes impacts using the overall student-level variation.

Table 1. Baseline equivalence before matching

Variables	Treatment			Control			Mean Difference	Effect Size (Hedges' *g*)
	N	Mean	SD	N	Mean	SD		
Prior Math (SY2324)*	214	3595.94	58.92	261,536	3576.15	59.42	19.80	0.333
Hispanic %	122	57.01	49.62	125,499	47.99	49.96	9.02	0.181
American Indian %	6	2.80	16.55	16,935	6.48	24.61	-3.68	-0.149
Asian %	6	2.80	16.55	15,178	5.80	23.38	-3.00	-0.128
African American %	23	10.75	31.04	24,345	9.31	29.06	1.44	0.050
White %	183	85.51	35.28	216,570	82.81	37.73	2.70	0.072
Native Hawaiian %	5	2.34	15.14	3,234	1.24	11.05	1.10	0.099
ELL %	16	7.48	26.36	25,093	9.59	29.45	-2.11	-0.072
SWD %	25	11.68	32.20	36,851	14.09	34.79	-2.40	-0.069
Migrant %	0	0.00	0.00	1,626	0.62	7.86	-0.62	-0.079
Female %	97	45.33	49.90	128,370	49.08	49.99	-3.75	-0.075

*Matching used grade-centered prior scores, but we report the raw scale scores and WWC-aligned mean differences.

Table 2. Baseline equivalence after matching

Variables	Treatment			Control			Mean Difference	Effect Size (Hedges' *g*)
	N	Mean	SD	N	Mean	SD		
Prior Math (SY23-24)*	214	3595.94	58.92	214	3595.71	59.37	0.23	0.004
Hispanic %	122	57.01	49.62	122	57.01	49.62	0.00	0.000
American Indian %	6	2.80	16.55	6	2.80	16.55	0.00	0.000
Asian %	6	2.80	16.55	6	2.80	16.55	0.00	0.000
African American %	23	10.75	31.04	23	10.75	31.04	0.00	0.000
White %	183	85.51	35.28	184	85.98	34.80	-0.47	-0.013
Native Hawaiian %	5	2.34	15.14	5	2.34	15.14	0.00	0.000
ELL %	16	7.48	26.36	15	7.01	25.59	-0.47	0.018
SWD %	25	11.68	32.20	26	12.15	32.75	-0.47	-0.014
Migrant %	0	0.00	0.00	0	0	0.00	0.00	0.000
Female %	97	45.33	49.90	97	45.33	49.90	0.00	0.000

*Matching used grade-centered prior scores, but we report the raw scale scores and WWC-aligned mean differences.

Impact model

We analyzed data using a two-level Hierarchical Linear Model (HLM) to account for the nested structure of the data, where students (level 1) are nested within schools (level 2). The student-

This does not impact the significance level; however, it may result in a more conservative estimate of the effect size. In this report, we followed the WWC recommendation and reported the effect size using student-level pooled variance.

level fixed effect covariates included students' prior year score (grade-mean centered), treatment status, and grade level. Grade 4 served as a reference grade in the model.

$$AASA_{ij} = \gamma_0 + \beta_1(PriorAASA_{ij}) + \beta_2(Treatment_{ij}) + \beta_3(Grade_{ij}) + \mu_{0j} + e_{ij}$$

Where:

- $AASA_{ij}$ = SY24-25 AASA mathematics scale score for student i in school j
- $PriorAASA_{ij}$ = grade-mean centered SY23-24 AASA mathematics scale score for student i in school j
- Treatment = dichotomous indicator of Navvy usage for student i in school j (0 = did not use Navvy at all, 1 = used Navvy for more than 65% of standards)
- Grade = grade level of student i in school j, with Grade 4 as the reference category.

SAS 9.4 (SAS Institute Inc, 2023) was used to estimate the models with the restricted maximum likelihood estimation method.

Results

Tables 3 and 4 present the number of students and schools, as well as the scale score descriptive statistics, for the matched sample by grade, respectively. The correlations between the grade-mean-centered prior score and the current year scale score ranged from 0.78 to 0.82.

Table 3. Descriptive statistics for the matched sample

Grade	Treatment		Comparison	
	Student	School	Student	School
4	52	4	52	30
5	12	1	12	11
6	8	1	8	6
7	29	1	29	12
8	113	4	113	38
Total	214	9	214	90

Table 4. SY2024-25 AASA scale score descriptives for the matched sample

Grade	Treatment				Comparison			
	Mean	SD	Min	Max	Mean	SD	Min	Max
4	3560.37	44.05	3466	3645	3547.08	44.10	3448	3645
5	3587.17	33.77	3541	3640	3580.58	41.95	3506	3640
6	3622.25	27.55	3572	3646	3601.38	52.36	3521	3676
7	3627.34	31.38	3577	3688	3630.34	33.40	3582	3704
8	3671.61	37.76	3602	3776	3654.82	34.07	3595	3769
Total	3632.00	60.27	3466	3776	3619.16	58.82	3448	3769

We calculated Intraclass Correlation Coefficients (ICCs) to quantify clustering in our outcome data and validate our analytical approach. ICCs indicate the extent to which variance in student outcomes is attributable to schools, which directly impacts the precision of treatment effects. Higher ICCs suggest students within the same school perform more similarly to each other, necessitating multilevel modeling to avoid biased standard errors. The ICC_{null} was 0.60, indicating students in the same school were highly similar to each other and supporting the two-level HLM modeling approach.

Research Question 1: Impact of Navvy usage on mathematics achievement

Table 5 presents results from our primary analyses, which include all grades in a single model with grade included as a fixed effect. This approach allows us to estimate the overall treatment effect while accounting for systematic differences in achievement across grades that are inherent to the vertical scale. Results show that after adjusting for school-level differences, treatment students scored on average 14 scale score points higher than comparison students within the same grade, corresponding to a moderate effect size of 0.24.

Table 5. Primary overall impact model

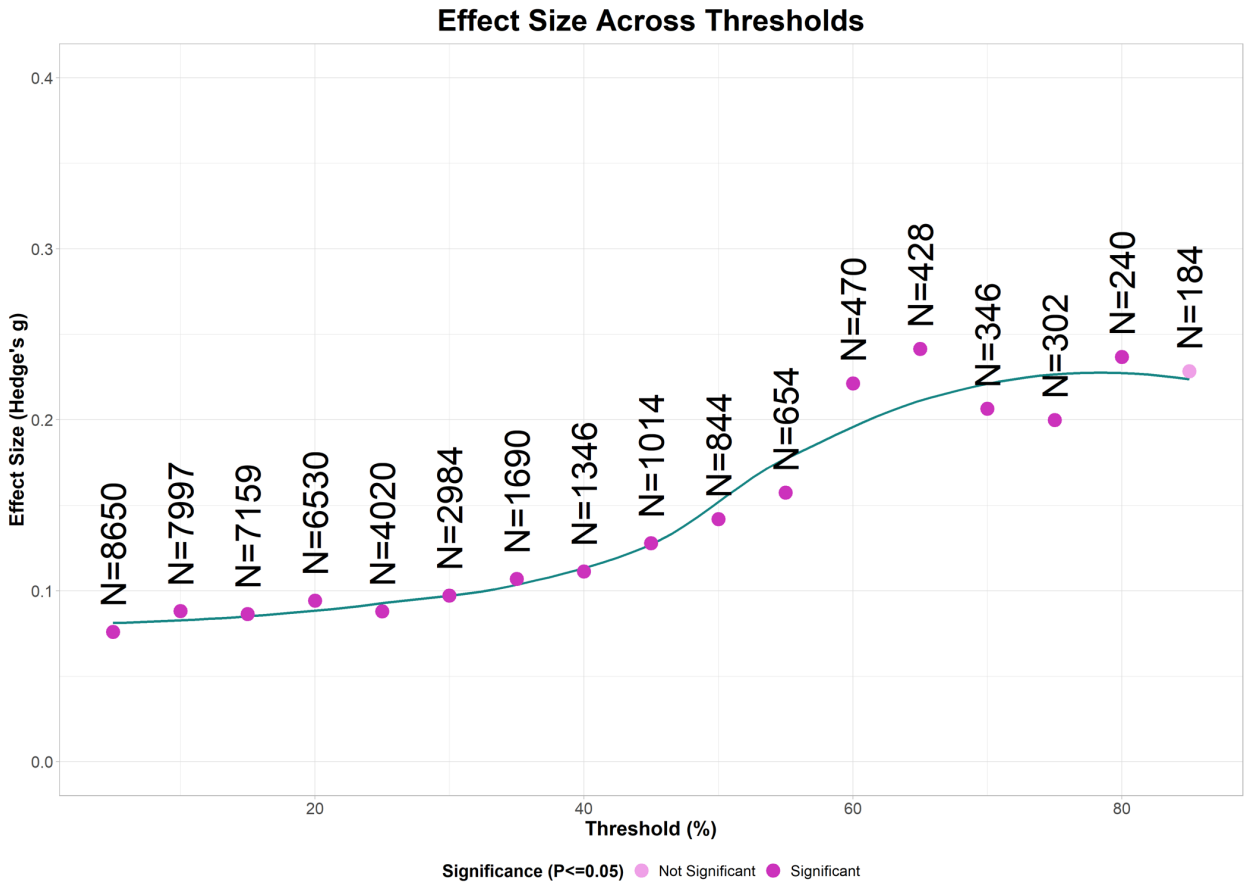
Effect	Estimate	SE	DF	t Value	Pr > |t|	Hedge's g
Intercept	3544.50	3.481	95	1018.14	<.0001	
Prior score	0.85	0.030	326	28.20	<.0001	
Treatment	14.41	4.960	326	2.91	0.004	0.24
Grade 5	32.71	7.257	326	4.51	<.0001	
Grade 6	67.30	8.239	326	8.17	<.0001	
Grade 7	78.12	5.161	326	15.14	<.0001	
Grade 8	107.08	4.249	326	25.20	<.0001	

Research Question 2: Relationship between Navvy participation and mathematics achievement

To address the second research question, "Among students who use Navvy, what is the relationship between the level of participation and their mathematics achievement?", we employed a dosage-like response analysis using varying participation thresholds. We established discrete cut points representing the percentage of standards in which students participated through Navvy assessments, matched respective comparison groups, and then estimated treatment effects at each threshold.

Figure 1 presents the estimated treatment effects and corresponding effect sizes (Hedges' g) at each participation threshold. The analysis reveals that the optimal treatment effect occurs when students participate in Navvy assessments for at least 65% of mathematics standards, with diminishing returns observed beyond this threshold. Students who participated in Navvy assessments for at least 60% of mathematics standards scored, on average, 13.48 points higher on the Arizona state mathematics assessment compared to similar students who did not participate in Navvy. The effect size of 0.22 indicates a meaningful impact on student mathematics achievement. Notably, the treatment effect at the 85% threshold was not statistically significant ($p < .05$), which may be due to the relatively small sample size at this level of participation. Appendix A shows the full statistics table.

Figure 1. Effect size with N counts at various participation rate



References

Ho, D., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*, 1-28.

SAS Institute Inc. (2013). *SAS® 9.4* (Computer software). SAS Institute Inc.

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/handbooks

Appendix A. Summary of Participation Effect at Various Participation Rate

Cut (%)	Total N	Estimate	Standard Error	t-value	P-value	Hedge's g
05	10770	4.31	1.09	3.95	<.0001	0.08
10	9464	4.97	1.18	4.21	<.0001	0.09
15	7788	4.86	1.30	3.74	0.00	0.09
20	6530	5.31	1.40	3.81	0.00	0.09
25	4020	4.91	1.50	3.27	0.00	0.09
30	2984	5.50	1.69	3.26	0.00	0.10
35	1690	6.13	2.02	3.04	0.00	0.11
40	1346	6.40	2.49	2.57	0.01	0.11
45	1014	7.45	2.89	2.58	0.01	0.13
50	844	8.44	3.07	2.75	0.01	0.14
55	654	9.71	3.95	2.46	0.01	0.16
60	470	13.48	4.68	2.88	0.00	0.22
65	428	14.41	4.96	2.91	0.00	0.24
70	346	12.00	4.52	2.65	0.01	0.21
75	302	11.23	4.89	2.29	0.02	0.20
80	240	12.25	5.17	2.37	0.02	0.24
85	184	11.96	6.07	1.97	0.05	0.23