Newark Study White Paper

Key points

- This white-paper describes the results of a three-year, longitudinal study of elementary and middle school students in a high-poverty, urban, U.S. school district. The purpose of the study was to evaluate the causal effect of Khan Academy on student achievement.
- Students who increased their usage of Khan Academy year over year saw a corresponding improvement in their achievement growth.
- Trends were consistent across student demographic subgroups, such as race/ethnicity, gender, and eligibility for free/reduced-price lunch.

Introduction

Newark North Ward (hereafter referred to as the district) is part of the Newark Public School District. In the 2023-2024 school year the 13 schools in the North Ward taught over 6000 students in grades 3-8. The student population is predominantly Hispanic/Latino and Black or African American. Over 90% of the students participate in the Free and Reduced Price Lunch program.

In November of 2022-2023 school year the district partnered with Khan Academy to leverage Khan Academy Districts (KAD) offering as a supplement to its curriculum. Prior to that some schools in the district used a free version of Khan Academy in their math classrooms. In the 2023-2024 school year, all 13 schools in the district implemented KAD for the entire school year.

The analysis described in this report assessed the impact of KAD usage on the performance on the New Jersey Student Learning Assessments (NJSLA). NJSLA is a statewide, computer-based test administered annually to students in grades 3-11 in New Jersey to measure their progress in English Language Arts (ELA), mathematics, and science according to the state's learning standards. In this report we focus on the mathematics NJSLA score.

Across grades and courses NJSLA mathematics scale scores span 650 to 850 on a vertically equated scale. Consequently, a scale score indicates the same latent proficiency irrespective of grade or test form; e.g., 750 represents "Meets Expectations" at any grade level. Performance levels are defined as: Level 1 (<700, "Did Not Meet"), Level 2 (700 - 724, "Partially Met"), Level 3 (725 - 749, "Approaching"), Level 4 (750 - 786, "Meets"), and Level 5 (≥787, "Exceeds").

The analytic window covers three academic years (2021-2022 through 2023-2024) and employs quasi-experimental methods to estimate the causal effect of KAD (operationalized as learning hours and skills leveled up to proficient or higher (STP+)) on subsequent NJSLA performance. Section 2 details the 2023-2024 analytic sample; Section 3 outlines the longitudinal panel construction and statistical methodology.

Study Sample

Table 1 below shows the student distribution by grade and school in the 23-24 school year. The majority (87%) of the nearly 6700 total students in grades 3-8 were on free or reduced priced lunch. Forty eight percent of the students are male, 77% are Hispanic or Latino, and 13% are Black or African American. Roughly 30% of the students were English language learners (ELL).

Table 1: *Overall Sample Size*

		Grade					
School	n	3	4	5	6	7	8
Abington Avenue	561	81	92	91	102	100	95
Dr. E. Alma Flagg	326	44	51	50	65	56	60
Dr. William H. Horton	453	55	65	62	91	86	94
Elliott Street	522	68	84	81	96	99	94
First Avenue	688	118	110	114	120	114	112
Franklin	469	62	73	104	84	76	70
Luis Munoz Marin	404	62	65	57	71	79	70
McKinley	387	51	71	63	62	69	71
Park	468	63	78	77	92	77	81
Rafael Hernandez	312	45	60	58	39	52	58
Ridge Street	360	52	57	62	59	58	72
Roberto Clemente	460	67	75	79	73	88	78
Salome Urena	271	44	54	50	49	48	26

Table 2 summarizes student achievement by NJSLA mathematics performance level in 2023-2024. Empirically, the district-level year-to-year mean gain on the vertically scaled NJSLA mathematics assessment is 2 - 4 scale-score points ($\mu \approx 3.4, \sigma \approx 21$) - insufficient to shift most students across the 25-point performance-level boundaries. As a result, categorical proficiency status is fairly stable across years (see Table 3).

Table 2: *Overall Sample NJSLA Performance*

		% at level			
Year	Mean	1	2	3	4
21-22	715.4	30%	35%	23%	12%
22-23	719.0	27%	32%	25%	16%
23-24	721.7	25%	32%	24%	18%

Note: Levels 4 and 5 collapsed for brevity.

Table 3 shows the proportion of students who moved up or down at least one NJSLA performance level in 2023-2024 relative to 2022-2023.

Table 3: *NJSLA Level changes by grade*

Grade	n	NJSLAΔ	Inc.	Dec.
Overall	4,241	2.7	1,126 (27%)	867 (20%)
4	816	3.4	222 (27%)	139 (17%)
5	835	3.6	213 (26%)	163 (20%)
6	879	0.2	195 (22%)	202 (23%)
7	905	9.2	315 (35%)	97 (11%)
8	806	-3.4	181 (22%)	266 (33%)

Note: Inc./Dec. indicates students who increased or decreased an NJSLA level from the previous year. Grade 3 excluded because that is the first year they take the test.

Khan Academy Usage

The district relied on Khan Academy illustrative math courses as a supplement during the core math instruction. Over the last two years the average time devoted to practice on Khan Academy increased from 20 hours to 26 hours over the school year. That is equivalent to 25 and 40 minutes per week respectively. Khan academy suggests at least 30 minutes per week of practice time based on its prior efficacy studies. In the 22-23 school year roughly 50% of the students met that level of usage. In the 23-24 school year that number increased to over 70% of students.

While learning time is a useful indicator of engagement, mastery of the skills on Khan Academy is often a more precise measure of learning on Khan. Reaching mastery on specific skills on Khan Academy indicates that a student actually worked through practice problems (as opposed to watching videos for example) and solved these problems correctly. If a student is using their time effectively, they can level up to proficient or mastered status on 2 or more skills per week or about 60+ skills over the course of the school year. In the 22-23 school year about 30% of the students leveled up on 60+ skills, whereas over 50% of the students reached that mark in the 23-24 school year.

The student sample characteristics in 23-24 school year are consistent with those observed in years 22-23 and 21-22.

Methodology

In observational settings, treatment take-up is endogenous: higher-achieving or more motivated students and their teachers are more likely to engage with supplemental tools such as Khan Academy. Naïve comparisons of users vs. non-users are therefore biased by both observed and unobserved covariates (e.g., motivation, teacher quality, family support).

Traditional matching strategies (e.g., exact, nearest-neighbor, or propensity-score matching) mitigate bias only to the extent that all relevant covariates are observed. Unmeasured attributes remain uncontrolled, matched samples shrink, and extreme propensity weights can inflate variance of the estimates. Typically only baseline test scores, grade, school and teacher are observed and used for matching and those are rarely sufficient to equate users and non-users.

An alternative approach is to use panel designs, where students are observed at multiple points in time. The identification of the causal effect of treatment comes from student level variation in treatment from time t to time $t+1$. For example a student who used Khan academy for 15 minutes per week in year 1 and then increased his usage by 10m per week on average in year 2 would be expected to increase their test scores in year 2 as compared to year 1. Under this design there are two sources of confounding - non-time varying confounding factors and time-varying confounding. Non-time varying factors include fixed student and teacher characteristics like overall student ability and teacher classroom preferences that dont change over time. Time-varying confounding comes from factors that change over time along with the changes in treatment levels. For example, if a school were to adopt a new curriculum in year 2 that included an increase in treatment as well as introduction of other interventions designed to increase math achievement, then the effect of the treatment would be confounded by the effects of the other interventions.

Fixed effects regression with student and teacher fixed effects

In our study we generate a panel data set with three years of student observations. The data set restricts students who are observed in at least any of the two years during the 3-year period. The students included in the analysis are not restricted in terms of their usage, but the students who did not use Khan Academy in all years in which they were observed are dropped from the analysis by construction of the analytical technique. The panel data set contains 5200 students. We used the fixed effects least squares regression approach to account for the non-time varying confounding described above. This approach works by comparing each student to themselves over multiple points in time, after subtracting each students overall average treatment and outcome values from each observation - a student fixed effect. This process (known as "demeaning") effectively removes factors that do not change over time. Similarly, if teachers are observed over multiple years, we can remove the teacher average test scores or treatment use by the same process.

This approach estimates the effect of within-student changes in treatment (Khan Academy use) on within-student changes in outcomes, effectively controlling for all stable student and teacher characteristics. However it does not account for time-varying confounding. Our model accounts for time-varying shocks common to all students by including a year fixed effect, but cannot remove shocks that are correlated with both dosage and outcomes at the school or classroom level (e.g., a concurrent adoption of a new math curriculum in year t that also increases Khan usage). Where data permit, we will test robustness by adding school-by-year and teacher-by-year fixed effects, which could remove time-varying confounding due to school-level policy or classroom-level changes, but the efficacy of this approach depends on the availability of data.

Analysis and Results

Most of the students increased their usage by 1-20 hours over the school year. About 10% increased their usage by 35 or more hours, and about 25% of the students decreased their usage.

The effect of Khan Academy use on student achievement is estimated using the following two-way fixed effects model

$$Y_{ijgt} = \beta T_{ijgt} + \alpha_i + \lambda_j + \delta_g + \gamma_t + \varepsilon_{ijgt} \tag{1}$$

where α_i is a student fixed effect, λ_j is a teacher fixed effect, δ_g is a grade fixed effect, γ_t is a year fixed effect, and ε_{ijgt} is the error term. The coefficient β represents the mean within-student change in standardized NJSLA points associated with a one-unit increase in Khan Academy usage (measured as one STP+ or one hour of learning time), conditional on the included fixed effects. All coefficient estimates are reported with heteroskedasticity-robust 95% confidence intervals.

Table 4 shows the estimated effect of Khan Academy usage on NJSLA test scores. Standard errors are clustered at the student and teacher level; 95% confidence intervals (CI) are shown in brackets.

Table 4: Fixed effect regression results

	(1)	(2)
Learning Hrs.	0.006*** [0.005, 0.008]	
STP+		0.003*** [0.003, 0.004]
Num.Obs.	12 942	12 942
R2	0.883	0.885
R2 Adj.	0.794	0.796
R2 Within	0.013	0.022
R2 Within Adj.	0.013	0.022
RMSE	0.34	0.34
Std.Errors	by: id & teacherId	by: id & teacherId
FE: id	X	X
FE: teacherId	X	X
FE: grade	X	X
FE: year	X	X

* p < 0.05, ** p < 0.01, *** p < 0.001

Average treatment effect

After controlling for year, grade, student, and teacher fixed effects, we estimate that each additional hour of Khan Academy practice improves NJSLA test score by +.006SD [.005,.008], and each additional skill to proficient improves the NJSLA test score by +.003SD [.003,.004]. Over 32 weeks in a school year this would translate to an average gain of +.10SD [.008, .123][1] from increasing Khan Academy practice by 30m per week, and an average gain of +.19SD [.16,.22][2] from increasing skills to proficient by 60 skills.

In practical terms these effects translate to about 2-4 additional NJSLA point gain from increasing practice time by 30m per week (18 hours over the school year) or 5-8 additional NJSLA points from increasing skill mastery by 2 skills per week (or 60 skills over the school year).

Heterogeneous treatment effects

We examine heterogeneous treatment for demographic subgroups and NJSLA baseline test score, which is the first available NJSLA test score depending on the year in which we first observed the student. Figure 1 shows the estimated effects for each subgroup after controlling for the same fixed effects as in the original model. Whereas the effects are consistent across the demographic subgroups like gender, ethnicity and ELL, the effect for students who at baseline were scoring below expectations (<700 on NJSLA scale) is statistically significantly larger. This suggests that weaker students may benefit more. Time-varying confounders correlated with both dosage and subgroup membership (e.g., remedial interventions targeted at low scorers) could upward-bias the baseline-ability interaction. For example, if low scoring students are concentrated in schools which encouraged both Khan usage and implemented other interventions, this effect will be spuriously inflated.

To probe this possibility we re-estimated the model with

[1] .0062*16 [.0047*16,.0077*16]
[2] .0031*60 [.0026*60, .0031*60]

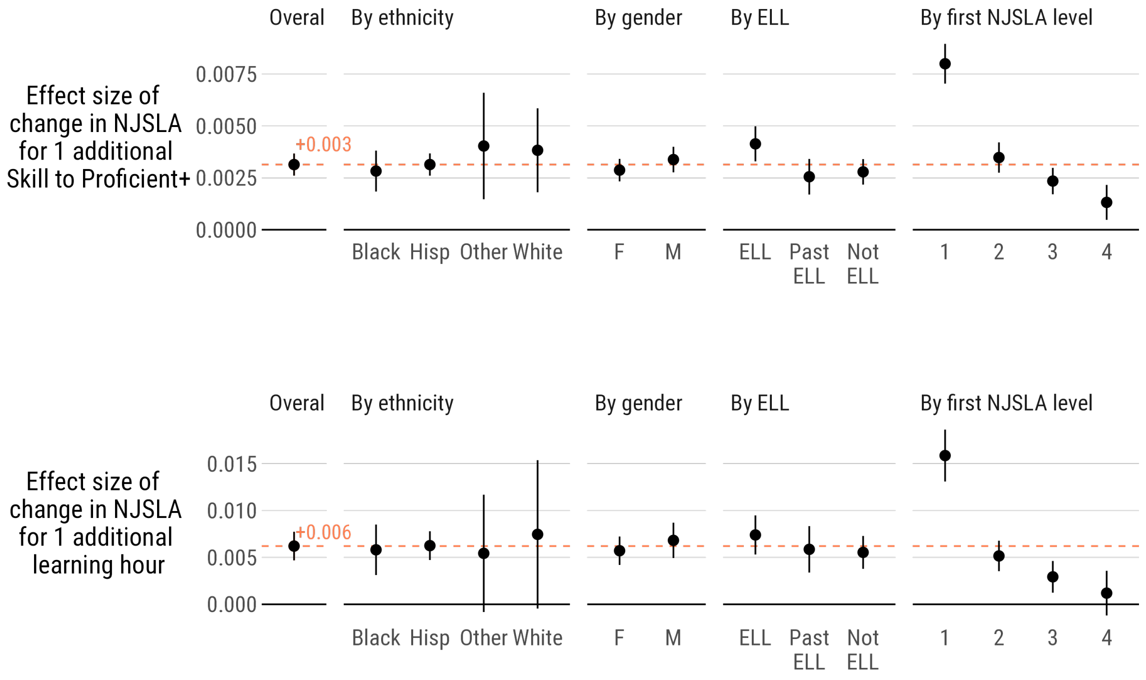

Figure 1: Summary of heterogeneous treatment effects

school-by-year fixed effects and teacher-by-year fixed effects. These terms would absorb any policy or program that is common to all students in the same school during the same academic year (e.g., a building-wide tutoring initiative aimed at low scorers) or efforts by individual teachers to improve test scores in their classrooms. After including these fixed effects, the baseline-ability by dosage interaction remained significant. This could have several explanations: 1) teachers could use individual student pull-outs for lower scoring students, or individual tutoring could be occurring and is correlated with both test scores and Khan usage; 2) low-scoring students could in fact benefit more from 1 hour of practice or 1 skill; 3) low scoring students could spend their time in a more beneficial way (e.g., focusing on prerequisites). The last two explanations are perhaps less likely than targeted student tutoring. However, given that for strong students (baseline 4) the effect of learning time is practically 0 whereas the effect of an STP is not, suggests that there may be a more effective way to spend time on Khan.

Additional robustness tests

We conducted two additional checks of our model. First we included a lagged dosage $X_{i,t-1}$ to conduct the lagged-effect test in order to rule out that the beta for treatment captures decaying effects of prior usage. Second we conducted a placebo test by including subsequent year dosage $X_{i,t+1}$ and estimating its effect on the current year NJSLA scores. A significant term for subsequent year dosage would imply that current year performance impacts usage decisions in a future year, which would violate the exogeneity assumption. To conduct these tests we restricted the sample to the students observed in all three years (n=2,924). Table 5 shows the coefficients from these two models. The lagged effect is statistically significant, suggesting that there is some carryover from the prior year. Given that it is about a third of the treatment effect and is estimated net of the main effect, it does not drive the main result and simply indicates that students recover some small benefit from prior year practice. The coefficient for subsequent year dosage is not statistically significant, strict-exogeneity has not been contradicted.

Table 5: Robustness test results

	(1)	(2)
Learning Hrs.	0.008*** [0.005, 0.010]	0.007*** [0.004, 0.009]
Learning Hrs. (lag)	0.003* [0.001, 0.005]	
Learning Hrs. (lead)		−0.002 [−0.005, 0.001]
Num.Obs.	5701	5468
R2	0.911	0.916
R2 Adj.	0.803	0.802
R2 Within	0.014	0.016
R2 Within Adj.	0.013	0.015
RMSE	0.30	0.28
Std.Errors	by: id & teacherId	by: id & teacherId
FE: id	X	X
FE: teacherId	X	X
FE: grade	X	X
FE: year	X	X

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Collectively, the results support a modest but educationally meaningful causal impact of sustained Khan Academy engagement on NJSLA mathematics achievement, with amplified benefits for students who entered the panel substantially below proficiency.

Conclusions

Leveraging a three-year, studentteacher two-way fixed-effects panel of 5,500 Newark North Ward students, we find that increased engagement with Khan Academy is associated with statistically and practically meaningful gains on the NJSLA mathematics assessment.

Specifically, dosage effect of an additional hour of practice is 0.006 SD increase in scale score; 30 minutes/week (18 hour/year corresponds to 0.10 SD (4 points). Mastery effect of each skill level-up is 0.003 SD; reaching 60 mastered skills in a school year produces 0.18 SD (7 points). Benefits are uniform across gender, ethnicity, and ELL status, but significantly larger for students who began below the proficiency cut-score (<650).

These results imply that Khan Academy usage can double the districts typical annual NJSLA growth. However the following limitations remain: residual time-varying confounding at the classroom and student level cannot be fully ruled out; always-zero users exit the analytic sample by design; and effects were measured only for mathematics.

Even with these caveats, the evidence indicates that sustained, mastery-focused use of Khan Academy can be an effective component of middle-grade mathematics instruction in high-poverty urban districts.