Efficacy of Online Learning Assessment Tools for COVID-19 Recovery Intervention

GEORGIA
POLICY LABS

Appendix to:

The Efficacy of Online Learning Assessment Tools for COVID-19 Recovery Intervention

Jennifer Darling-Aduana and K. Jurée Capers

Metro Atlanta Policy Lab for Education

November 2024

Technical Analysis Specifications

We used ordinary least squares (OLS) regression, school fixed effects, and inverse probability-weighted regression adjustment (IPWRA) models to examine the relationship between any i-Ready usage and achievement test score growth. We also estimated models that added an indicator for whether any lessons were teacher-assigned, which are manually assigned to students by their teachers—versus automatically assigned lessons based on the adaptive learning platform. We then used dose-response functions to examine differential associations with achievement growth by number of i-Ready lessons. Lastly, we conducted an analysis of heterogeneous associations by running each model limiting the sample to subgroups of interest.

Ordinary Least Squares Regression

To estimate the extent to which student use of online instructional materials was associated with improved student achievement growth in the post-pandemic era, we estimated an OLS regression model comparing student outcomes accounting for i-Ready access using data from SY 2021-22 and SY 2022-23 (Equation 1). Data were limited to kindergarten through Grade 9, since no i-Ready use was observed in Grades 10 through 12.

$$\mathrm{score}_{\mathrm{igt}} = B_0 + B_1 \mathrm{iReady}_{\mathrm{it}} + B_2 \mathrm{days}_{\mathrm{it}} + B_3\, \mathrm{score}\, e_{\mathrm{it}-1} + X_{\mathrm{it}} B + \theta_g + \sigma_t + \varepsilon_{\mathrm{igt}} \quad (1)$$

We estimated separate models for math and reading scores. All models control for the number of days between the beginning- and end-of-semester test, lagged test scores from tests administered at the beginning of the school year, a vector of individual-level student characteristics (X_{it}), grade fixed effects (θ_g), and year fixed effects (σ_t). Student characteristics include student race/ethnicity and gender, as well as FRPM eligibility, English Learner (EL) status, and identified disability status (separated into identified emotional, intellectual, or physical disability). We also accounted for whether the student received any disciplinary referrals. All models used robust standard errors clustered at the school level.

Next, given the district policy that provided earlier i-Ready access to some schools over others, we estimated the same model shown in Equation 1, adding school fixed effects (λ_s)—as shown in Equation 2.

$$score_{isgt} = B_{0+} B_1 iReady_{it} + B_2 days_{it} + B_3 score_{it,n-1} + X_{it} B + \lambda_s + \theta_g + \sigma_t + \varepsilon_{isgt} \quad (2)$$

We estimated Equation 2 twice: once with all cases and once limiting data to cases on common support from the propensity score analysis. Removing cases off common support has the benefit of preventing outliers without viable comparison cases from skewing estimates but reduces sample size and limits generalizability. More information on the proportion of cases removed due to this specification is provided when discussing IPWRA modeling below.

Inverse Probability-Weighted Regression Adjustment Analysis

When implementing the OLS regression model, concerns remain related to treatment and comparison group imbalance, which can introduce bias into estimates.[1] To minimize this concern, we also employed IPWRA to control for the extent to which there were systematic differences in whether a student used i-Ready based on both pre-treatment and fixed characteristics. The principal benefit of IPWRA includes balancing the treatment and comparison groups on a vector of relevant observable variables, which often results in dropping cases off common support that are so extreme that there are no comparable cases.

To implement this approach, we first estimated the likelihood of receiving treatment (i.e., using i-Ready or having a teacher-assigned lesson) using logistic regression conditioned on the vector of baseline school-level covariates (X_{it}) included in Equation 1 (see also Appendix Table A1) related to the prior achievement, socio-demographic characteristics, grade, year, and school CSI/Horizon status (see Equation 3).[2]

$$\Pr(z_i=1|X_i)=\frac{1}{(1+\exp[-(X\beta)])} \quad (3)$$

Propensity scores were calculated using matching with a 0.01 caliper, no replacement, and limited to cases on common support. This method resulted in an 81–82% reduction in bias when estimating any i-Ready use and a 77–79% reduction in bias when estimating any teacher-assigned i-Ready lessons (see Appendix Table A2). Approximately 5% of the analytic sample was dropped when estimating any i-Ready use due to being outside the range of common support. The sample was more balanced when estimating whether a student had any teacher-assigned i-Ready lessons, which resulted in dropping less than 0.01% of cases.

Appendix Table A1. Predicting Treatment Variables from Matching Covariates

	Treatment variables			
	Any math i-Ready	Any reading i-Ready	Ever teacher-assigned (math)	Ever teacher-assigned (reading)
Fall same-subject MAP	-0.05***	-0.04***	0.05***	0.01
Inst. days between tests	0.01***	0.01***	0.02***	0.02***
Race/ethnicity (other baseline)				
Black	0.12***	0.20***	0.42***	0.42***
White	-0.61***	-0.59***	-0.87***	-0.96***
Hispanic	-0.66***	-0.48***	-0.65***	-0.41**
Female	0.04***	0.07***	0.07***	0.07***
Experiencing homelessness	-0.38***	-0.39***	-0.20*	-0.25**
FRPM	0.13***	0.15***	0.05**	0.06**
Identified disability	-0.16***	-0.14***	-0.25***	-0.26***
English Learner	0.40***	0.50***	0.32***	-0.03
Grade (K baseline)				
1	0.11***	0.10***	0.59***	0.73***
2	0.21***	0.24***	1.02***	1.19***
3	0.27***	0.36***	1.35***	1.42***
4	0.21***	0.31***	1.59***	1.36***
5	0.13	0.24***	1.43***	1.36***
6	-0.42***	-0.32***	-0.55***	0.01
7	-0.88***	-0.71***	-0.07	-0.19***
8	-1.01***	-0.85***	0.19***	0.33***
9	-8.65***	-8.61***	N/A	N/A
Year (2021–22 baseline)				
2022–23	0.21***	0.36***	0.06**	0.31***
Horizon school	1.08***	1.19***	0.88***	1.00***
Ever CSI	1.22***	1.12***	0.58***	0.52***
Number of cases	105,970	96,280	95,809	86,612

Appendix Table A2. Propensity Score Matching Reduction in the Bias

		Off support	LR chi2	Mean bias	% bias	Variance Ratio
Any math i-Ready	Unmatched	–	12,045.74	27.6	86.3	1.96
	Matched	5,221	512.13	5.2	23.5	1.38
Any read. i-Ready	Unmatched	–	10,614.72	27.0	84.0	2.04
	Matched	5,263	418.83	4.9	21.9	1.48
Ever teacher-assigned (math)	Unmatched	–	7,040.38	23.4	92.9	0.78
	Matched	72	176.26	3.5	18.8	1.08
Ever teacher-assigned (read.)	Unmatched	–	5,761.78	23.6	87.1	0.93
	Matched	2	186.94	3.2	20.3	1.01

The likelihood ratio chi-squared test results show a substantial decrease in differences between groups on matching variables. The mean bias column reports the difference in means between the treatment and comparison group after accounting for the overall spread of each group. For each treatment variable, the mean bias between groups decreased dramatically. The matched samples had mean bias ranging from 4.7 to 9.7, compared to 16.8 to 27.6 for the unmatched sample. The estimated percent bias remaining in the sample similarly decreased between the unmatched and matched samples. The matched models estimating any i-Ready use (in math and reading), as well as whether the teacher ever assigned a lesson in reading, fell below the 25.0 rule of thumb, while the matched model estimating whether the teacher ever assigned a lesson in math was approaching that threshold at 36.5% bias. For well-balanced groups, it is also desirable to have a variance ratio between 0.5 and 2, with a ratio of 1 being ideal. The matched models estimating any i-Ready (in math and reading) as well as whether the teacher ever assigned a lesson in math fell within this range. The matched model estimating whether a teacher ever assigned an i-Ready lesson in reading showed a small improvement in the variance ratio but remained below the 0.5 threshold. Lastly, we identified substantial overlap in propensity scores between treatment and control groups visually, with the weighted propensity score distributions contributing to additional improvements in the distribution of propensity scores between groups, which is required to minimize bias.[3]

After estimating the likelihood of receiving each treatment, we used the resulting propensity score to fit a weighted regression model with covariate adjustment. This model was identical to the OLS regression equation presented in Equation 1, apart from the inclusion of the weights generated from the matching process described above. More specifically, when assigning weights, treatment participants received a weight of 1/(ρ)^, while control participants received a weight of 1/(1-ρ)^.[4] The advantage of this doubly robust method is that the estimates will be unbiased if either the propensity score or regression adjusted models are correctly specified.

To examine our second research question, we estimated separate OLS and IPWRA (Equations 1–3) that accounted for whether any teacher assigned an i-Ready lesson as the treatment in addition to any i-Ready use. If a lesson was not teacher-assigned, then the lesson was assigned by the automated i-Ready system.

Dose-Response Function

For the second research question, we also estimated dose-response functions to estimate the effect of different doses (or amount of exposure) to i-Ready; this approach is appropriate for use when assignment to treatment is non-random and accommodates continuous treatment measures.[5] For these functions, we consider the number of i-Ready lessons assigned as the continuous treatment variable. We estimate two functions: one where the dose is zero (Equation 4) and one where the dose is not zero and continuous (Equation 5).

$$w = 0 \rightarrow y_0 = B_0 + B_1 days_{it} + B_2 score_{it,n-1} + X_{it}B + \varepsilon_{it} \tag{4}$$

$$w = 1 \rightarrow y_1 = B_0 + B_2 days_{it} + B_3 score_{it,n-1} + X_{it}B + h(t) + \varepsilon_{it} \tag{5}$$

Equation 5 adds to Equation 4 the term *h(t)*, which is a response function that accounts for the dose (i.e., how many i-Ready lessons the student was assigned). We subsequently estimate and graph the resulting non-linear response curves. Given that the modal number of lessons assigned were 20–25 (depending on the subject) with standard deviations of 20–24, we chose to graph only up to 100 lessons to prevent attempts to interpret estimates where there was insufficient data to support robust findings. This choice resulted in dropping no more than 1% of the sample.

As a sensitivity check we generated graphs including these outliers; maximum average treatment effects and trends were comparable whether dropping or including outliers. As with previous analyses, results are associational in nature. We also examined correlations between the number of assigned i-Ready lessons and fall MAP test scores to determine whether students might be systematically assigned more or fewer i-Ready lessons based on their prior achievement level, which could bias the resulting estimates. Among students assigned any i-Ready lessons, we observed no significant correlation between the number of math lessons assigned and students' fall MAP math score ($r = 0.01$, $p = 0.27$). We observed a significant, negative correlation in reading ($r = -0.05$, $p < 0.01$). This translates into being assigned one fewer i-Ready reading lesson for each standard deviation increase in MAP reading test scores. Although significant, this is a low-magnitude relationship given that, on average, i-Ready users were assigned 20 reading lessons. For this reason, we are less concerned that systematic assignment of i-Ready lessons by prior achievement level might be biasing the resulting estimates.

Heterogeneity Analysis

Lastly, we examined potential heterogeneous treatment effects by estimating Equations 1–3, limited to each student subgroup of interest (i.e., gender and race/ethnicity as well as EL status, FRPM eligibility, identified disability, and disciplinary status). We also examined, separately, students who did not meet state standards on the prior year state level and grade level (lower elementary, upper elementary, middle, and high school). Additionally, we estimated the dose-response function analysis for students with various prior levels of achievement and by FRPM status.

Limitations

Before proceeding to our findings, we reiterate the limitations of our empirical strategy. Findings are based on data from one metro-Atlanta-area district, which has its own unique contextual and demographic factors that may not generalize to all settings. Our analyses are descriptive versus causal as we cannot fully account for endogenous variation.[6] Relatedly, we were unable to fully understand or account for selection into treatment. Among the approximately 67,000 instances where peers in a students' school-grade-year were assigned to use i-Ready, only around 28,000 students were assigned lessons. Given low universality, we estimated models that examined whether the students' school-grade-year had access to i-Ready interacted with whether that student used i-Ready. Estimates for any use of i-Ready remained comparable in directionality, magnitude, and significance, minimizing our concern that not being able to fully account for selection into treatment within school-grade-year bands might be biasing our results.

Importantly, we were unable to collect data on or account for other recovery interventions occurring during the same period that might have been targeting similar schools. And, lastly, we conducted an intent-to-treat (ITT) analysis that examined all students assigned to complete any i-Ready lesson (versus those that actually completed an i-Ready lesson). ITT is an inherently conservative approach and thus may underestimate the true benefit for students who are adequately supported and motivated to fully engage. However, we felt conducting this analysis was appropriate to inform district decision-making. We attempt to provide insight into the possible maximum benefit for students who engage fully with the i-Ready platform by estimating dose-response functions with average treatment effects across the number of i-Ready lessons assigned.

Endnotes

1. We also attempted a two-stage least squares (2SLS) approach. However, we do not report these estimates as our excluded variables (i.e., Horizon and CSI status) were not valid based on post-estimation tests. This indicated that Horizon and CSI status were associated with the outcome independent of assignment to treatment and should be included in the second-stage model. IPWRA was preferable because, as a doubly robust estimator, we could include these indicators both in the original matching models and in the second-stage regression models.

2. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55

3. Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

4. Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5-86

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

5. Cerulli, G. (2012). *A continuous treatment model for estimating a dose response function under endogeneity and heterogeneous response to observable confounders: Description and implementation via the State module "ctreatreg."* Rome, Italy: National Research Council of Italy.

6. As an attempt to account for characteristics that are fixed over time (i.e., innate intelligence, parental education level), we implemented a student fixed effect approach to compare within-student changes in test scores between years when a student used i-Ready versus years when the student did not use i-Ready. However, these models could only be estimated on a substantially reduced sample (14% of the full sample) and violated the strict exogeneity assumption, likely due to the differential effects the COVID-19 pandemic had on student achievement based on characteristics associated with socioeconomic status, race, and ethnicity, and the possibility of regression to the mean after near universal lower-than-expected achievement growth during the pandemic. Even despite these limitations, estimates were comparable in direction and magnitude, although not always significant.