Working Paper

Methods Used to Estimate Achievement Effects in Personalized Learning Schools

John F. Pane and Matthew Baird

RAND Education

WR-1061-BMGF
November 2014
Prepared for the Bill & Melinda Gates Foundation

Methods Used to Estimate Achievement Effects in Personalized Learning Schools

John Pane and Matthew Baird, RAND Corporation

The purpose of this document is to describe the methods RAND used to analyze achievement for 23 personalized learning (PL) schools for the 2012-13 through 2013-14 academic years. This work was performed at the request of the Bill & Melinda Gates Foundation (BMGF), as part of a multi-year evaluation contract. The 23 schools were selected from a larger portfolio of PL schools funded directly or indirectly by BMGF because they implemented PL school-wide during both of the two academic years and they also administered Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) both years.

The MAP assessment enables the comparison of achievement of students in the PL schools (the treatment group) with students in the same grade in other schools (the virtual comparison group, or VCG) who are similar in terms of achievement and the demographic characteristics of their schools.

Matching Criteria for Virtual Comparison Groups

For each treatment student, NWEA created a VCG of up to 51 students from its database. Separate comparison groups were created for the mathematics and reading tests and for each time span examined. The following student and school matching criteria were applied to create the VCG:

Requirements for All VCG Matches

- Students have valid scores from fall 2012 and spring 2014.
- Students are in different school districts from the treatment group students.
- Schools have the same locale classification (e.g., urban, suburban, rural, etc., according to the National Center for Educational Statistics (NCES) Public School Universe Survey).
- Students are the same gender and in the same grade as the treatment group students to whom they are matched.

Approximate Matching Criteria[1]

- Schools differ by no more than 15 percentage points on the portion of students participating in the national free or reduced-price lunch program (FRL).
- Students scored within 5 points on NWEA's RIT scale on the fall 2012 MAP assessment.
- Days elapsed between fall 2012 and spring 2014 testing differs by no more than 18 days.

[1] NWEA first identifies all student records that meet these criteria, and then takes a random sample of 51 of these records.

Assessment of balance between the treatment group and the VCG

The VCG is intended to be very similar to the study group in terms of students' observable characteristics prior to treatment. This is true by construction for the criteria that were matched exactly (namely, the grade level of the student and the urbanicity of their school). For the approximate matching criteria, we examined whether the groups appear to be the same, controlling for the grouping of each study student with up to 51 VCG students (on average, PL students were matched to 48 VCG students). Table 1 shows balance on variables that were approximately matched. In both mathematics and reading, very close matches were achieved on the starting MAP scores. The school percentages of students eligible for FRL were 2-3 percent higher in the VCG. The number of days between the fall 2012 and spring 2014 assessments were about six days longer for the PL students than for the VCG students. We use the doubly robust technique of including each of these variables as covariates in outcomes models.

Table 1: Balance between PL and VCG Groups on matching variables

		VCG mean	PL mean	Group difference	Standardized group difference
Mathematics	Starting test score (RIT)	179.25	179.36	0.11	0.00
	School FRL percentage	89.09	86.35	-2.74	-0.29
	Elapsed time between tests (days)	625.00	631.03	6.03	0.31
Reading	Starting test score (RIT)	175.53	175.56	0.03	0.00
	School FRL percentage	88.09	85.90	-2.19	-0.22
	Elapsed time between tests (days)	623.78	629.66	5.88	0.29

Note: Group difference is the difference in means between the PL and VCG groups; the standardized group difference is the weighted average of differences within each stratum (a PL student and the student's VCG), divided by the standard deviation of the pooled sample. Standardized differences were calculated using the R function xBalance from package RItools.

Analytic methods

To analyze the effect of attending a PL school, we fit hierarchical linear models that account for clustering of students within schools and of each student with his or her virtual comparison group of up to 51 students. The dependent variable in this model is the gain from fall 2012 to spring 2014 in the MAP assessment scale score (i.e., spring 2014 RIT score minus fall 2012 RIT score). We standardize test scores using mean and standard deviations of the fall 2012 RIT scores by grade, so that the fall scores have a mean of zero and standard deviation of one within each grade level, and spring scores reflect the standardized growth. Model coefficients can thus be read as standardized gains in spring scores. The regression models include as covariates the percentage of students eligible for FRL, and the elapsed time (in days) between fall and spring assessments. Given the within-VCG estimation strategy, none of the exactly-matched covariates can be included in the regression, but are implicitly controlled for by the estimation strategy.

For comparison, we also estimate the treatment effect using conditional expected growth estimates based on norms (CGN) provided for each student by NWEA (for more on their methodology, see Northwest Evaluation Association (2011). For each relevant subgroup (school, grade span, or overall), we estimate the average difference between the realized treated student growth and the conditional growth estimates for if they were not PL schools.

In order to estimate relative rates of growth (either overall or for a subsample, such as a school or grade span), we estimate the average ratio of the expected growth if treated and the expected growth if untreated, for all treated students in the relevant sample. The expectations are estimated by taking the estimated coefficients from the hierarchical linear models and, the observed characteristics of each student, predicting their growth under treatment and without.[2]

The result of this calculation can be interpreted (taking the example of school-by-school estimates) as a growth rate for each school in the study relative to a hypothetical school with an observably similar student body in terms of grade level, gender, and baseline test scores, as well as school-level urbanicity and FRL eligibility.

Sensitivity test

Many of the PL schools are schools of choice, where families make an affirmative decision to enroll their children. Family involvement in education might influence student achievement in positive ways unrelated the schools' influence on achievement. To the extent VCGs are drawn from schools that are not schools of choice, there is the potential that a difference in family involvement could bias the results. We investigated this concern by attempting to make the treatment and control groups more similar in terms of the family involvement implied by enrollment in schools of choice. We requested NWEA to create an additional VCG, supplementing the matching criteria with the additional restriction of drawing only from other schools of choice, which NWEA defined as charter, academy, private, magnet, and parochial schools. The results from the schools of choice VCG are very similar to the preferred results. In particular, there is not evidence the preferred results are meaningfully biased by using the standard VCG matching criteria that ignore choice.

Adjustments for multiple hypothesis tests

Because these analyses involve numerous tests of statistical significance on the same dataset, we applied the Benjamini & Hochberg (1995) False Discovery Rate method to adjust each test's threshold for statistical significance. We grouped 56 mathematics effect size and relative growth rate estimates into one domain, and 56

[2] We also test estimating the denominator (expected growth given no treatment) by simply taking the average growth of that student's VCG scores. This yields very similar results, which is to be expected, given the regression coefficients are based on the growth of the VCG and treated students.

reading effect size and relative growth rate estimates into a second domain and applied the Benjamini-Hochberg algorithm to each domain with the target for post-correction alpha=0.05.

Limitations

Although the analysis of MAP using VCGs is the most rigorous method available, it relies on the matching of students within our sample to similar students outside the sample. Because students in treatment schools may differ from their comparison groups in unobserved ways that affect their academic performance, this method is vulnerable to selection bias even if matches appear to be very good on observable characteristics. Any unmeasured differences between the study sample and the comparison group can result in biased estimates of treatment effects. Also, this analysis implicitly assumes that VCG students are in more traditional schools that are generally not implementing PL innovations. There is no way to verify this assumption. To the extent PL is actually more widespread in VCG schools, the contrast between treatment and VCG instruction would be reduced and the analysis would underestimate the effects of PL. Because of these limitations, achievement results should be interpreted with some caution.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289-300.

NWEA. (2011). RIT Scale Norms: For Use with Measures of Academic Progress® (MAP®) and MAP for Primary Grades. Portland, OR: Northwest Evaluation Association.