Confidential, NOT FOR DISTRIBUTION or ATTRIBUTION

evaluation engine

Enabling quasi-experimental impact evaluations using state longitudinal data

TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed

for TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed

Generated by Evaluation Engine on 4/17/2017

1. OVERALL RESULTS

This report shows differences between participants in the intervention of interest (TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed) and a comparison group of students with similar or identical background characteristics and prior academic achievement. Information about participants and control group students is drawn from the state's longitudinal data system (SLDS).

The SLDS contained suitable matches for 1820 program participants. Each of these 1820 participants was matched to one, two, three or four non-participants, with a total of 6994 distinct non-participants serving as matched controls.

The impact estimates for TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed are presented in the figures and tables that follow. The figures show the combined estimates (comparisons of averages over all matched participants and over all matched comparison values), while the tables show subgroup-specific estimates (if requested).

OUTCOME: Algebra I: High school end-of-course exam, 2013

Participant outcomes typically fell below outcomes of matched counterparts

There was an error generating this chart.

Figure 3: EOC exam-taking rate, Algebra I

Data: Taken and passed Algebra 1 (end-of-course) exam, year 2 after intervention start. (0 = Did not take exam, 1 = Took exam but did not pass, 2 = Passed exam.) The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes fell below the matched control group's. This difference is unlikely to be due to chance (p<0.025, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 260 standard deviations, if it is a strong predictor of the outcome, or by 2000 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Attendance, 2011

Participant outcomes typically fell below outcomes of matched counterparts

Effect Size:
0.00

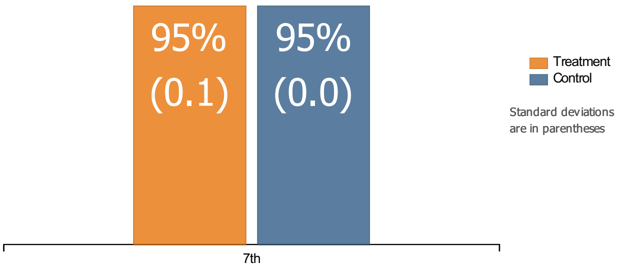


Figure 4: Attendance rates

Data: Attendance rate for academic year of intervention start.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes fell below the matched control group's by 0 percentage points, on average. This difference is unlikely to be due to chance (p<0.025, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 4.5 standard deviations, if it is a strong predictor of the outcome, or by 34 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Attendance, 2012

Participant outcomes typically fell below outcomes of matched counterparts

Effect Size:
-0.01

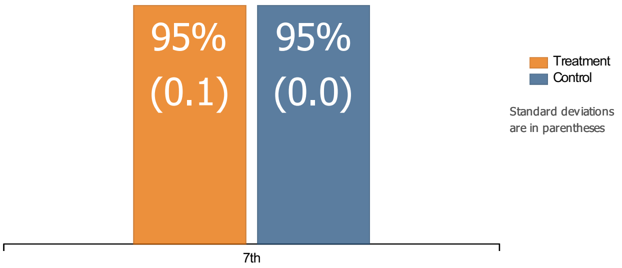


Figure 5: Attendance rates

Data: Attendance rate for year 1 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes fell below the matched control group's by 0 percentage points, on average. This difference is unlikely to be due to chance (p<0.025, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 36 standard deviations, if it is a strong predictor of the outcome, or by 270 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Attendance, 2013

Participant outcomes typically exceeded outcomes of matched counterparts

Effect Size:
0.02

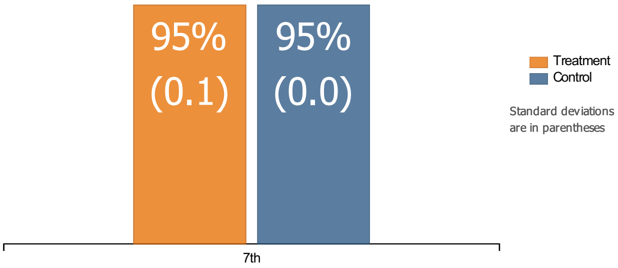


Figure 6: Attendance rates

Data: Attendance rate for year 2 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes exceeded the matched comparison group's by 0.1 percentage points, on average. This difference is unlikely to be due to chance (p<0.025, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must exceed those of their matched counterparts by at least 120 standard deviations, if it is a strong predictor of the outcome, or by 930 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: English II: High school end-of-course exam, 2013

Participant outcomes typically fell below outcomes of matched counterparts

There was an error generating this chart.

Figure 9: EOC exam-taking rate, English II

Data: Taken and passed English 2 (end-of-course) exam, year 2 after intervention start. (0 = Did not take exam, 1 = Took exam but did not pass, 2 = Passed exam.) The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes fell below the matched control group's. This difference is unlikely to be due to chance (p<0.025, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 800 standard deviations, if it is a strong predictor of the outcome, or by 6000 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Mathematics Achievement (TCAP), 2011

Participant outcomes typically fell below outcomes of matched counterparts

Effect Size:
-0.02

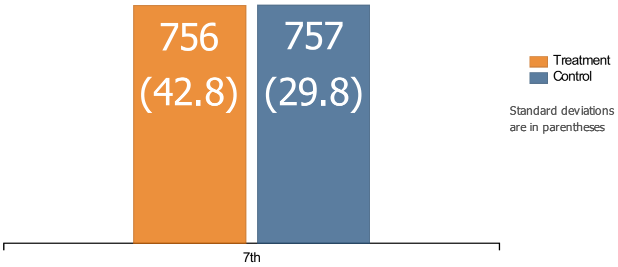


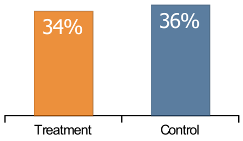
Figure 10A: Average scale scores in math

Data: Score on grade-level Math exam, academic year of intervention start.

Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Please see note below Figure 10B.

Overall Proficiency:



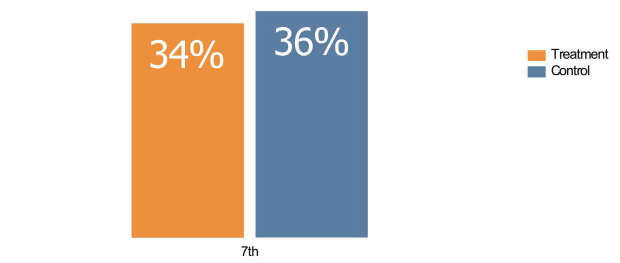


Figure 10B: Proficiency rates in math

Data: Score on grade-level Math exam, academic year of intervention start.

Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Notes: The intervention group's outcomes fell below the matched control group's by 0 scale score points, on average, corresponding to the difference between the 44.8th and 44.8th statewide percentiles. This difference is unlikely to be due to chance (p<0.025, one-sided, based on comparison of scale scores), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 45 standard deviations, if it is a strong predictor of the outcome, or by 340 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Mathematics Achievement (TCAP), 2012

Participant outcomes typically fell below outcomes of matched counterparts

Effect Size:
-0.01

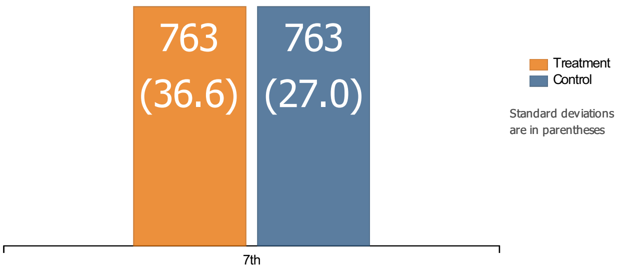


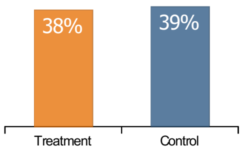
Figure 11A: Average scale scores in math

Data: Score on grade-level Math exam, year 1 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Please see note below Figure 11B.

Overall Proficiency:



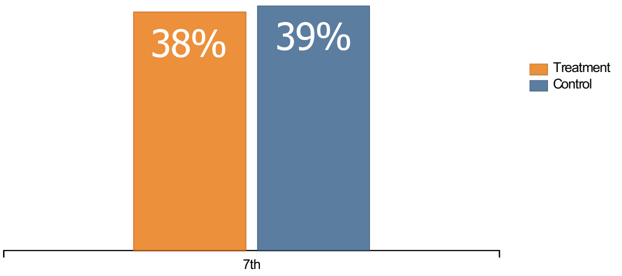


Figure 11B: Proficiency rates in math

Data: Score on grade-level Math exam, year 1 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Notes: The intervention group's outcomes fell below the matched control group's by 0 scale score points, on average, corresponding to the difference between the 47th and 47th statewide percentiles. This difference is unlikely to be due to chance (p<0.025, one-sided, based on comparison of scale scores), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 71 standard deviations, if it is a strong predictor of the outcome, or by 530 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Persistence, 2011

Participant outcomes typically fell below outcomes of matched counterparts

Effect Size:
0.00

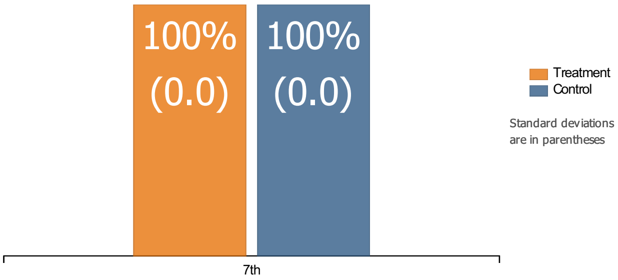


Figure 13: Persistence rates

Data: Persistence in academic year of intervention start.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes fell below the matched control group's by 0 percentage points, on average. This difference is unlikely to be due to chance (p<0.025, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 260 standard deviations, if it is a strong predictor of the outcome, or by 2000 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Persistence, 2012

Participant outcomes typically fell below outcomes of matched counterparts

Effect Size:
0.00

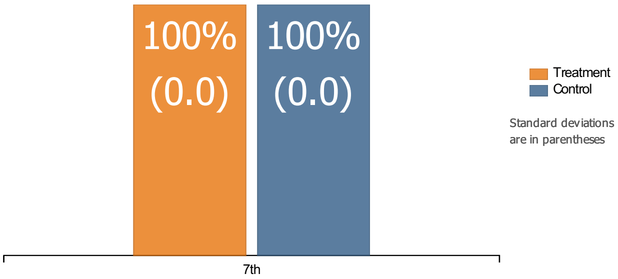


Figure 14: Persistence rates

Data: Persistence in year 1 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes fell below the matched control group's by 0.1 percentage points, on average. This difference is unlikely to be due to chance (p<0.025, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 410 standard deviations, if it is a strong predictor of the outcome, or by 3100 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Persistence, 2013

Participant outcomes typically fell below outcomes of matched counterparts

Effect Size:
0.00

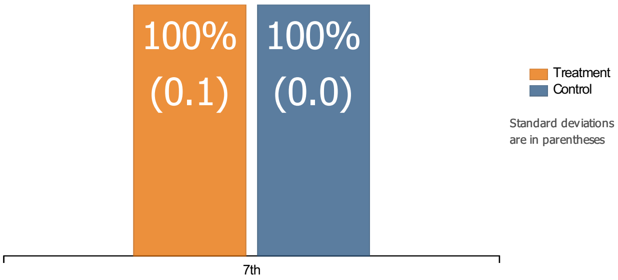


Figure 15: Persistence rates

Data: Persistence in year 2 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes fell below the matched control group's by 0.3 percentage points, on average. This difference is unlikely to be due to chance (p<0.025, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 880 standard deviations, if it is a strong predictor of the outcome, or by 6600 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Reading Achievement (TCAP), 2011

Participant outcomes typically fell below outcomes of matched counterparts

Effect Size:
0.01

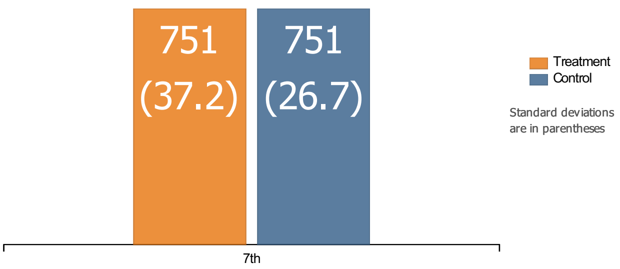


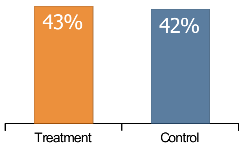
Figure 16A: Average scale scores in reading

Data: Score on grade-level Reading exam, academic year of intervention start.

Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Please see note below Figure 16B.

Overall Proficiency:



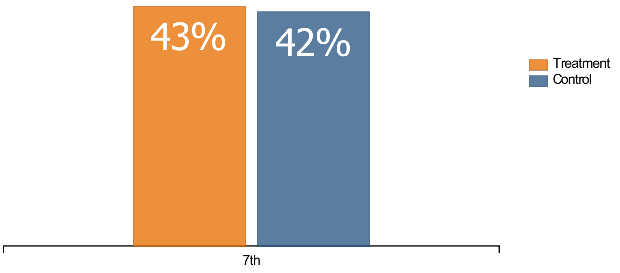


Figure 16B: Proficiency rates in reading

Data: Score on grade-level Reading exam, academic year of intervention start.

Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Notes: The intervention group's outcomes fell below the matched control group's by 0 scale score points, on average, corresponding to the difference between the 43.6th and 43.6th statewide percentiles. This difference is unlikely to be due to chance (p<0.025, one-sided, based on comparison of scale scores), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 12 standard deviations, if it is a strong predictor of the outcome, or by 88 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Reading Achievement (TCAP), 2012

Participant outcomes typically fell below outcomes of matched counterparts

Effect Size:
-0.01

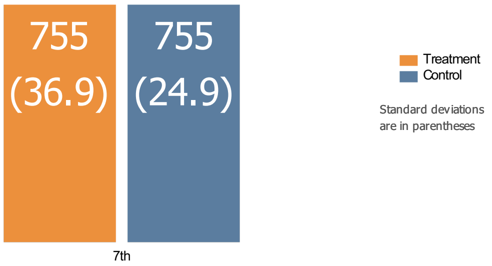
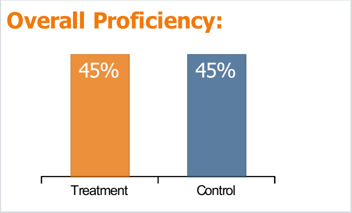


Figure 17A: Average scale scores in reading

Data: Score on grade-level Reading exam, year 1 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Please see note below Figure 17B.



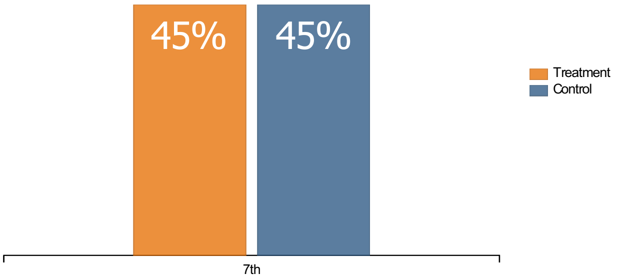


Figure 17B: Proficiency rates in reading

Data: Score on grade-level Reading exam, year 1 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Notes: The intervention group's outcomes fell below the matched control group's by 0 scale score points, on average, corresponding to the difference between the 42.8th and 42.8th statewide percentiles. This difference is unlikely to be due to chance ($p<0.025$, one-sided, based on comparison of scale scores), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 73 standard deviations, if it is a strong predictor of the outcome, or by 550 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Promotion, Current Year, 2011

Participant outcomes typically fell below outcomes of matched counterparts

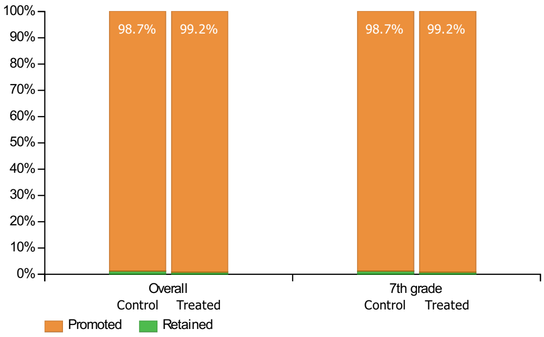


Figure 19: Promotion rates

Data: Promotion to next grade, academic year of intervention start.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes fell below the matched control group's. This difference is unlikely to be due to chance ($p<0.025$, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 460 standard deviations, if it is a strong predictor of the outcome, or by 3400 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Promotion, Current Year, 2012

Participant outcomes typically fell below outcomes of matched counterparts

There was an error generating this chart.

Figure 20: Promotion rates

Data: Number of times promoted within 2 years of intervention start The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes fell below the matched control group's. This difference is unlikely to be due to chance (p<0.025, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 360 standard deviations, if it is a strong predictor of the outcome, or by 2700 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

OUTCOME: Promotion, Current Year, 2013

Participant outcomes typically fell below outcomes of matched counterparts

There was an error generating this chart.

Figure 21: Promotion rates

Data: Number of times promoted within 3 years of intervention start The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: The intervention group's outcomes fell below the matched control group's. This difference is unlikely to be due to chance (p<0.025, one-sided), although unmeasured differences between groups might explain it. (To create such a bias, participants' values of a hypothetical omitted variable must fall below those of their matched counterparts by at least 350 standard deviations, if it is a strong predictor of the outcome, or by 2700 or more standard deviations if it is a weaker predictor. See Appendix, Sensitivity to unmeasured variables.)

2. SUBGROUP ANALYSIS

OUTCOME: Attendance, 2011

Effect estimates by subgroup, with tests for presence of effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	0.00	0.00
Limited-English proficient, current year		
NA	-0.07	1.00
No	-0.01	1.00
Yes	0.37	0.00

Data: Attendance rate for academic year of intervention start.

Source: Tennessee State Longitudinal Data System

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: Attendance, 2012

Effect estimates by subgroup, with tests for presence of effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	-0.01	0.00
Limited-English proficient, current year		
NA	-0.04	1.00
No	-0.01	1.00
Yes	0.14	1.00

Data: Attendance rate for year 1 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: Attendance, 2013

Effect estimates by subgroup, with tests for differences from overall effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	0.02	0.00
Limited-English proficient, current year		
NA	0.07	0.89
No	0.02	1.00
Yes	0.28	0.31

Data: Attendance rate for year 2 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: English II: High school end-of-course exam, 2013

Effect estimates by subgroup, with tests for differences from overall effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	0.00	0.00
Limited-English proficient, current year		
NA	0.00	0.54
No	0.00	1.00

Yes	0.00	0.86

Data: Taken and passed English 2 (end-of-course) exam, year 2 after intervention start. (0 = Did not take exam, 1 = Took exam but did not pass, 2 = Passed exam.) The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.
Source: Tennessee State Longitudinal Data System

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: Mathematics Achievement (TCAP), 2011

Effect estimates by subgroup, with tests for presence of effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	-0.02	0.00
Limited-English proficient, current year		
NA	-0.36	1.00
No	-0.01	1.00
Yes	0.08	1.00

Data: Score on grade-level Math exam, academic year of intervention start.
Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: Mathematics Achievement (TCAP), 2012

Effect estimates by subgroup, with tests for presence of effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	-0.01	0.00
Limited-English proficient, current year		
NA	-0.61	1.00
No	-0.01	1.00
Yes	0.39	0.00

Data: Score on grade-level Math exam, year 1 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.
Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: Persistence, 2011

Effect estimates by subgroup, with tests for differences from overall effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	0.00	0.00
Limited-English proficient, current year		
NA	0.00	1.00
No	0.00	0.10
Yes	0.00	1.00

Data: Persistence in academic year of intervention start.
Source: Tennessee State Longitudinal Data System

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: Reading Achievement (TCAP), 2011

Effect estimates by subgroup, with tests for differences from overall effect

	ESTIMATED BENEFIT	

	(EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	0.01	0.00
Limited-English proficient, current year		
NA	-0.12	1.00
No	0.01	0.94
Yes	0.11	0.88

Data: Score on grade-level Reading exam, academic year of intervention start.

Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: Reading Achievement (TCAP), 2012

Effect estimates by subgroup, with tests for presence of effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	-0.01	0.00
Limited-English proficient, current year		
NA	-0.31	0.00
No	0.00	1.00
Yes	0.55	0.00

Data: Score on grade-level Reading exam, year 1 after intervention start. The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System, Tennessee Comprehensive Assessment Program (TCAP) Achievement

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: Promotion, Current Year, 2011

Effect estimates by subgroup, with tests for differences from overall effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	0.00	0.00
Limited-English proficient, current year		
NA	0.00	1.00
No	0.00	0.42
Yes	-0.02	1.00

Data: Promotion to next grade, academic year of intervention start.

Source: Tennessee State Longitudinal Data System

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: Promotion, Current Year, 2012

Effect estimates by subgroup, with tests for differences from overall effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	0.00	0.00
Limited-English proficient, current year		
NA	-0.02	1.00
No	0.01	0.41
Yes	-0.02	1.00

Data: Number of times promoted within 2 years of intervention start The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.

Source: Tennessee State Longitudinal Data System

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

OUTCOME: Promotion, Current Year, 2013

Effect estimates by subgroup, with tests for differences from overall effect

	ESTIMATED BENEFIT (EFFECT SIZE) WITHIN SUBGROUP	*P*-VALUE
Overall	0.01	0.00
Limited-English proficient, current year		
NA	-0.01	1.00
No	0.01	1.00
Yes	0.09	0.23

Data: Number of times promoted within 3 years of intervention start The grade level indicated in the figure margin is the students' grade level at the time the intervention began, not at the time the outcome was measured.
Source: Tennessee State Longitudinal Data System

Notes: Smaller p-values indicate less plausible null hypotheses. The uppermost p-value attaches to the null hypothesis of no effect or a negative effect. Since it is relatively small, the remaining rows present p-values for whether the intervention had similar effects for the specified subgroup as for participants overall.

3. DATA AND STATISTICAL METHODS

This part of the report describes the data used to evaluate the participant group (TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed), the analytic method used to assess differences in outcomes for the participant group compared with the matched control group of students not participating in the intervention and provides graphical and tabular information documenting the comparability of the participant and control groups.

Data

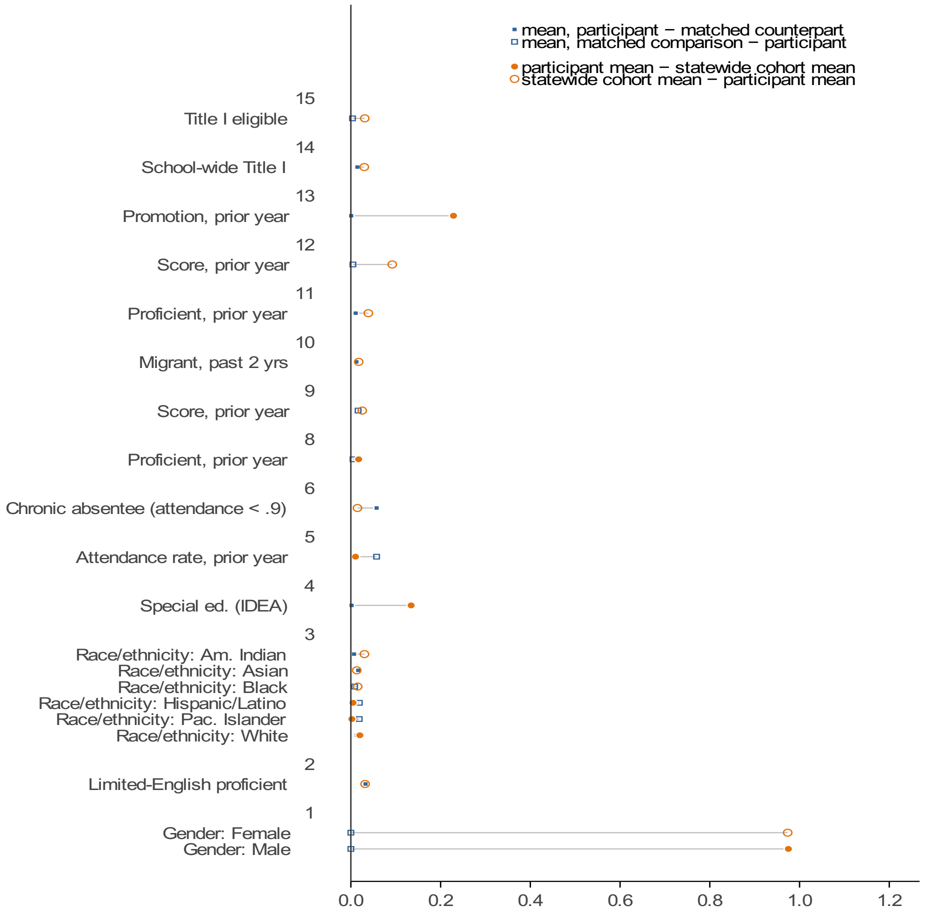
Student-level data for Tennessee for 2010-2012 were compiled by the Tennessee Consortium on Research, Evaluation, and Development (TNCRED) and provided by the Tennessee Department of Education’s Office of Research and Policy. School- and district-level characteristics were constructed by aggregating the student-level data. Additional school- and district-level information was obtained from the Common Core of Data maintained by the National Center for Education Statistics (NCES).

Measured covariates

The Evaluation Engine associates each program participant with 0-4 students who were in the same grade when the program began, according to the SLDS, but were not identified as belonging to the participant group. Although participants and their matched counterparts are required to have been enrolled at the same grade level when the intervention began, they are not required to be the same on any other specific pre-intervention characteristics. Instead, the Evaluation Engine assembles information from the SLDS into variables describing participants and eligible counterparts at the outset of the intervention— the *measured covariates*— before selecting matches using *propensity scores*[1] based on combinations of measured covariates. Characterized in terms of measured covariates, comparison groups assembled in this fashion ordinarily are near-equivalents of the participant groups they were matched to, although they may differ in terms of unmeasured baseline variables. Accordingly, Evaluation Engine propensity scores use an expansive list of measured covariates[2].

Balance Plot

Figure A1: Study group versus matched or statewide comparison group, at outset of the intervention



Notes: The figure compares the study group to students statewide at the same grade level, and to the matched comparison group, in terms of averages of their baseline characteristics. Horizontal positions of circular plotting symbols indicate the number of standard deviations by which study group means differed from state cohort means, whereas the x-axis positions of square plotting symbols indicate the number of standard deviations separating study group means from the matched counterpart means.

Measured covariates used in matching include familiar demographic and academic achievement variables, such as sex, race and economic disadvantage, and performance on prior years' achievement tests, are included, as are several baseline variables constructed for the Evaluation Engine, such as "smoothed" versions of prior years' test scores.[3] The main student-level variables used in the match are named in the left column of Appendix Figure A1.

The graphic that appears to the right of the variable names in Figure A1 shows the magnitudes of difference between participants in the intervention and other students in the state, along with the extent to which the matching process mitigated these differences. Specifically, absolute differences between the means for the intervention group and the means for all students in the same grade in the state are indicated with orange circles. The same differences for the participant group mean and the matched comparison group mean are indicated with blue squares. Since the different variables are measured in different units, each variable is standardized, that is, presented in multiples of the standard deviation of the variable in question. Thus, a variable having a symbol at 0.2 on the horizontal axis means that the participant group differed from the comparison group by 2/10 of a standard deviation.

When the intervention group's mean on a given variable was larger than that of the comparison group, the difference is represented with a solid (filled) circle or square; an unfilled symbol indicates that the comparison group's mean was larger. Figure A1 shows many but not all of the variables included in the matching process: besides a few more student variables, for example dummy variables for whether the student had a score on previous years grade level tests, the Evaluation Engine also adjusts for a number of school level variables, including school demographic and academic achievement profiles.

Figure A1 shows that after matching, differences between the two groups remain, even in terms of measured covariates. This is also true of randomized controlled trials (RCTs), the "gold-standard" method of estimating program impacts. In finite samples, random assignment leaves small, random differences between treatment and control groups. The Evaluation Engine aims to produce matched comparison groups that are about as similar to the participant group on measured characteristics as typically occurs in paired RCTs of the same size. This comparison to a hypothetical paired RCT can be made precise: in this analysis, if we could have randomly reassigned participation in the intervention between actual participants and their matched counterparts, the ex ante probability of obtaining baseline differences larger than those shown in Figure A1 would be 0%[4]. (In this instance, unfortunately, even the matched differences shown in the figure are strictly larger than the chance imbalances of random assignment. The Evaluation Engine addresses this possibility with post-matching covariate adjustments, described below.) When feasible, the matching procedure goes on to address imbalance at baseline on school- and student-level covariates in addition to those shown in Figure A1[5].

Propensity score matching aims to bring about a situation in which participant group means and matched comparison group means follow the same probability distributions, for measured covariates and perhaps other variables[6], just as they would in an RCT. Even when the match appears to have met its aims with regard to measured variables, the impact estimates it generates carry two important limitations.

First and more important: Although the Evaluation Engine may approximate an RCT in terms of measured covariates, it cannot replicate an RCT's ability to balance unmeasured baseline variables. If at baseline participants differ systematically even from non-participants who share their measured characteristics, and if these systematic differences in any way associate with outcomes, then the Evaluation Engine's impact estimates will be biased. Such differences are typical of programs for which students, classrooms or schools were deliberately selected, and selected on the basis of characteristics going beyond those recorded in the SLDS.[7]

Second, the matching procedure only emulates RCTs with *student-level* random assignment: for example, an RCT studying the effectiveness of a program by following the winners and losers of a lottery for seats in the program. As compared to designs involving random assignment of schools or classrooms, such RCTs may have difficulty documenting the effectiveness of programs which indirectly affect participants' classmates or schoolmates, even those who are not themselves program participants.[8] Similarly, the Evaluation Engine's approach is likely to lack power, by comparison with designs involving the matching of classrooms or schools, for programs of this type.

Matching procedure

Matches are made only within grade levels (at the start of the intervention). If the intervention group is particularly large, matches may also be restricted to fall within the levels of one or more of the subgroup analysis variables. Matching within these groups involves three propensity scores.

The first propensity score (PS) estimated models students' participation in the intervention as a function of demographic and achievement characteristics of their schools. Specifically, for each school the proportions of students falling in the various racial/ethnic, gender, economic disadvantage, English-language learner and special-education categories, shown in Figure A1, are tabulated. Additionally, racial/ethnic diversity is represented as the probability that two students selected at random from a given school would belong to the same racial/ethnic group.[9] Finally, the school-level propensity modeling procedure incorporates empirical Bayes estimates of school-level average achievement test results: in large schools, these averages fall very close to simple averages of student scores, whereas in smaller schools they "borrow strength" from other, demographically similar schools by adjusting the school average score closer to the average of those similar schools. In the school-level PS model, these school characteristics are used to characterize the likelihood of participation in TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed for this cohort. No matches are made on the basis of this first PS: rather, it is used to reduce the pool of controls eligible for matching to only those students attending schools not unlike schools attended by program participants.

The second propensity score that is fit models participation as a function of student characteristics, the core measures shown in Figure A1. The first match that the Evaluation Engine attempts is a match on the score from this model, termed the "core PS." This match is conducted to determine whether it is possible to balance the core student level variables between participants and matched controls. If an acceptable level of balance, defined as the level that random assignment would have achieved with probability .8, is attained for core student measures, the process proceeds to attempt matching on other characteristics. In difficult matching problems, however, the process stops after matching as closely as possible on the core propensity score only. In order to produce the best match, the matching routine permits participants to share matched control group students, if doing so allows for closer matches on the core PS; at the same time, students in the participation group may be matched to up to four controls.

If balance on core variables is found to be attainable, the algorithm attempts to match simultaneously on the core PS and on a third propensity score, the "inclusive PS," which predicts participation in the intervention based on student and school level variables concurrently. If by matching on a combination of the core PS and the inclusive PS the algorithm is able to balance both core measures and the inclusive list of covariates, then it explores whether the match can be modified to address several additional goals. First, it tries to match participation group students to comparable students in the same school districts. This is a back-up measure to ensure that matches come from similar school contexts[10], the primary measure being to match within the same state and to address school context variables in the school and in the inclusive PS. Second, it attempts to make matches within the same levels of a condensed disability categorization, to increase the likelihood that program participants taking alternate forms of state tests will be matched to non-participants taking the same alternate test.[11] Third, the procedure also aims to select matches for participants from schools in which few or no other students participated in the same intervention, to avoid matching students in the participation group to non-participants who may have been affected by the intervention indirectly, because of its presence at their schools -- (a "spillover" effect).[12] In this analysis, 99.8% of matches were made within district.

Given the varied aims of the matching procedure, it is inevitable that situations will arise in which there are simply no non-participants available who are in all respects good matches for a given participant. The Evaluation Engine addresses this in several ways. First, prior to matching, participation group students whose *inclusive* propensity scores differ measurably from those of each non-participant are removed from the analysis sample.[13] (A side-effect of this is to impose overlap, in terms of the inclusive PS, on the matched sample, by removing those participation group students whose propensity scores are very different from those of each non-participant.) Because the inclusive PS model coefficients have relatively large standard errors, participants are trimmed only if they are very different from the non-participant group: this is a relatively permissive overlap requirement.

Second, the Evaluation Engine procedure sets priorities among the different propensity scores, devoting attention to the inclusive propensity score only after confirming that matching on the main propensity score can closely balance the variables contributing to it, and devoting attention to within-district matching, within disability category matching and to drawing matches from lower treatment density schools only after it has succeeded in balancing inclusive PS variables as well. Matching well on all of these factors at once will not always be possible, but the Evaluation Engine ensures the best match, given what the data permits, by using provably optimal matching algorithms.[14]

Outcome analysis

Significance tests begin by comparing treatment and comparison subjects on a given outcome. The Evaluation Engine's method of comparison uses a Winsorization procedure to limit the influence of outlying observations; for achievement test outcomes at grades 4-8, it also incorporates a robust covariance adjustment for the prior year's test scores. These comparisons are then used in (normal-theory approximations to) permutation tests. Such tests assess the magnitude of an apparent trend in favor of the treatment group by repeatedly reassigning labels of students as treatment or control, independently in each matched set, and recording the apparent trend in favor of the notional treatment group that was created in this way; they then calculate the fraction of notional treatment groups enjoying advantages over the corresponding notional control groups as large or larger than the actual treatment group's advantage over its matched comparisons-the *p-value*. When this p-value is small, chance alone cannot readily explain the apparent trend in favor of the treatment group. A second, complementary p-value calculation characterizes the plausibility of a chance explanation for an apparent trend to the advantage of the control group. If either of these "one-sided" p-values falls below 2.5%, the difference between the groups is deemed statistically significant. When differences in scale scores and in proficiency levels are presented together, the significance level shown is for the scale score difference, not the proficiency level difference.

If subgroup analyses were requested, the significance levels that accompany them differ in meaning depending on whether the estimated overall effect was significant and in favor of the program. When participants' values of a given outcome are significantly higher than those of their matched counterparts, the subgroup analysis that follows tests whether the benefit experienced by the selected subgroups differed from the overall average benefit of participation in the intervention program. On the other hand, if participants' values tend to be lower than those of their matched counterparts (or if there is no consistent and statistically meaningful trend) then the Evaluation Engine tests for program benefits separately within each subgroup. The results of the two different types of subgroup analyses are presented in a similarly structured table.

Statistically significant main effect

In the case of a statistically significant main effect, covariances of main and subgroup effect estimates are estimated by an adaption of the method of the paired t test.[16] Because there can be many such tests, based on the number of categories in the subgroups, the chance of one or more spurious significant findings is high. To address this possibility, the Evaluation Engine incorporates a correction based the method of Hothorn *et al.*[15]

Non-significant main effect

In the event that this overall test is not statistically significant, then any subsequent subgroup analysis (if subgroup analysis was requested) is a separate permutation t-test for the presence of an intervention effect within each of the designated subgroups. Since the number of tests is large, the method of Hothorn *et al.* is used for multiplicity corrections.

Effect estimates

Estimates of the intervention effect are calculated with "effect of treatment on the treated" weighting. Specifically, for each student in the participant group matched to one or more non-participating students, the effect of the intervention is estimated as the difference between the participant's outcome and the average outcome among the participant's matched counterparts. The second of these numbers is called the participant's *matched counterpart value*. In simple cases, overall intervention impacts are estimated as averages of estimated effects on program participants—that is, as means of differences between participants' outcomes and corresponding matched counterpart values.

The calculation is more complex if there are participants who could not be matched, if outcome information was unavailable for one or more matched subjects or if some participants differed from their matched counterparts in terms of key demographic characteristics. If there are participants who could not be matched, they do not contribute to the estimate of the program's benefit. (The estimate is of the effect of treatment on *those treated subjects who could be matched*, not all treated subjects irrespective of matchability.) Participants matched only to controls lacking a value of a given variable do not contribute to the variable's participant mean, nor do these controls contribute to the variable's matched counterpart mean. Similarly, in order to contribute to the matched counterpart mean of a given variable, a non-participating student must be matched to a participant who is not missing a value on that variable. (Balance calculations apply similar rules when missing values are found in baseline variables.)

If some participants were matched to counterparts differing from them on sex, race/ethnicity, economic disadvantage, special education (IDEA), or limited English proficiency, participant minus matched comparison value differences are calculated on indicator variables encoding each demographic categorization. Then a linear model is used to adjust outcome differences, participants' outcomes minus corresponding matched comparison values, for these demographic covariate differences. (In contrast with models used for significance testing, this model is fit by ordinary rather than robust regression; it does not attempt to adjust for prior test scores.) The fitted model's intercept term, a covariance adjusted mean difference of participant and matched comparison outcomes, serves as the estimate of the program's benefit.

Sensitivity to unmeasured variables

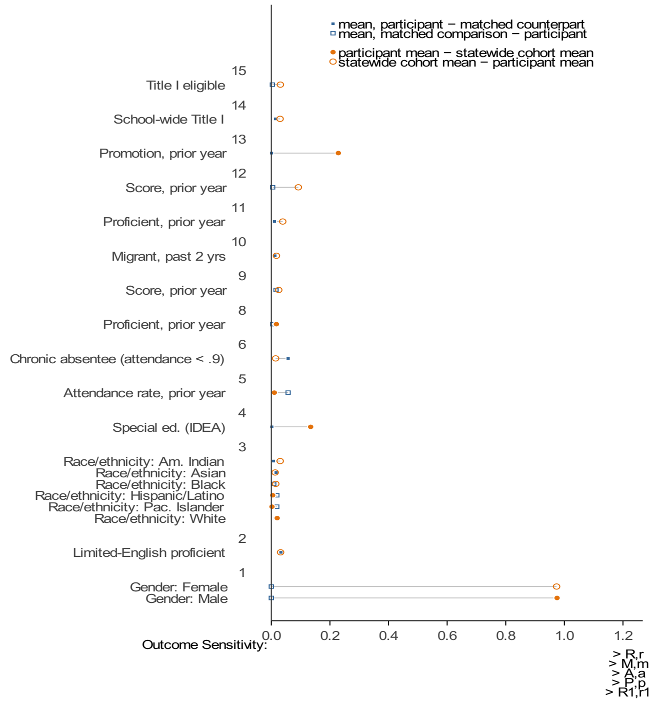
As a quasi-experimental method, the Evaluation Engine propensity match cannot ensure comparability of the groups in terms of unmeasured variables. Propensity score matching may help correct for some unmeasured variables, particularly those that are highly correlated with measured variables included in the matching; but unmeasured between-group differences can, if they are highly predictive of outcomes, greatly bias the Evaluation Engine's impact estimates. Ultimately it falls to the researcher to assess the potential for biases of this kind. Appendix Figure A2 offers some assistance in making these assessments.

If one or more outcome difference was found to be statistically significant, the extent of that finding's omitted variable sensitivity[17] is recorded beneath the x-axis of Figure A2. The markings that appear there indicate the minimum extent of the difference between students in the participation and matched control groups, in standard deviations of the hypothetical unmeasured variable, that would be necessary for that variable to explain away the impact finding (in the sense of rendering it statistically insignificant). Such an unmeasured variable could more readily upset a finding if it strongly predicted the relevant outcome than if its association with that outcome were weak. To represent the range of possibilities, therefore, for each statistically significant impact estimate the figure indicates two estimates of the extent of unmeasured difference necessary for spurious statistical significance: an uppercase letter, "R" for a reading achievement outcome, "M" for math and so forth, indicates the requisite imbalance for a hypothetical omitted variable that predicts the outcome strongly, having a partial correlation with that outcome of 0.75; whereas the horizontal position of the corresponding lowercase letter indicates the extent of confounding necessary for a weaker predictor of that outcome, one having a partial correlation with it of 0.1, to cause spurious statistical significance.

Partial correlations of .75 are quite high: prior achievement tests in the same subject as the posttest typically have partial correlations to the posttest of this magnitude; but few other baseline/outcome pairs of variables relate this strongly. Demographic variables' correlations with achievement outcomes are commonly about .1.[18]

To analyze the sensitivity of a significant finding to a specific omitted variable, situate the omitted variable between these extremes in terms of its likely relationship with the outcome, and then consider what degree of participant-matched control difference on the variable is plausible. In some cases, hypotheses about this difference are easier to formulate and understand by comparing them to measured variables. In such a case, one can calibrate sensitivity by comparing measured differences from the upper part of the plot to sensitivity thresholds on the bottom of the plot. For example, consider an unmeasured variable with a partial correlation to the grade level reading test outcome, say, of 0.1, with differences between the participant group and matched controls on par with these groups' pre-matching difference in eligibility for free or reduced-price lunch, for instance. Such an omitted variable could sufficiently bias an estimated grade level reading effect to make it falsely statistically significant only if the red circle for Free/reduced price lunch falls to the right of the small "r" symbol at the bottom of the plot. An omitted variable on which participants and matched controls are imbalanced to a degree similar to this, but which correlates more strongly with the reading outcome, could more readily produce spurious statistical significance. Specifically, for an unmeasured variable with partial correlation to the same reading outcome of .75, the appropriate reference point is the uppercase "R" at the bottom of the figure.

Figure A2: Sensitivity of outcome findings to omitted variable confounding (below x-axis) against measured baseline differences (above x-axis)



[Click here to open in new window]

Notes: The figure compares the study group to students statewide at the same grade level, and to the matched comparison group selected by the Evaluation Engine. Differences are expressed as a proportion of each variable's standard deviation. When estimated program impacts reach statistical significance, the magnitude of the difference in a hypothetical omitted variable that would be needed to upset the finding is indicated beneath the x axis.

[1] P.R. Rosenbaum and D.B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70:41–55, 1983.

[2] Indeed, an important reason the Evaluation Engine uses propensity score matching is the capability of the technique to address more baseline variables than other common methods of adjustment for quasi-experiments. See [Rubin and Thomas, 1996].

[3] This smoothing involves the use of so-called "Empirical Bayes" techniques to average students' observed test scores with mean scores at their schools and demographic groups. The smoothed scores permit meaningful substitute scores for students missing a baseline test result, and they limit bias from test measurement error in models using test scores as an explanatory variable. Because the smoothed scores are calculated using baseline information exclusively, they are pre-intervention variables, and matching on them does not contaminate the outcome comparison (D.B. Rubin, "For objective causal inference, design trumps analysis," *Annals of Applied Statistics*, 2(3):808–40, 2008).

[4] The Evaluation Engine's covariate balance summary follows Hansen, B.B. and Bowers, J., "Covariate balance in simple, stratified and clustered comparative studies," *Statistical Science*, 23(2):219–236, 2008.

[5]For this evaluation of TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed, post-matching balance on the full list of covariates, core covariates plus school-level variables and additional student measures, was at the 0 percentile of balance (on comparable collections of covariates and in comparable RCTs). See the Supplemental Statistical Material for additional detail.

[6] See Rosenbaum, P.R. and Rubin, D.B, "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70:41–55, 1983; Rubin, D.B., "Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism," *Biometrics*, 47:1213–1234, 1991; or Braitman, L.E. and Rosenbaum, P.R., "Rare Outcomes, Common Treatments: Analytic Strategies Using Propensity Scores," *Annals of Internal Medicine*, 137(8):693–695, 2002.

[7] Specific selection scenarios can sometimes be translated into plausible ranges of bias, and in turn into how much statistically adjusting for an unmeasured variable would have caused study findings to change, with the help of the lettered annotations in the bottom margin of Figure A1. See the discussion "Sensitivity to Unmeasured Variables," above.

[8] See e.g. Raudenbush, S.W., "Statistical analysis and optimal design for cluster randomized trials," *Psychological Methods*, 2:173–185, 1997; Sobel, M.E., "What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference," *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

[9] In other words, one of the independent variables contributing to the school level PS is the probability of interspecific encounter (Hurlbert, S.H., "The nonconcept of species diversity: a critique and alternative parameters" *Ecology*, 52(4):577–586, 1971) as applied to race/ethnicity. Lucas and Berends ("Sociodemographic diversity, correlated achievement, and de facto tracking," *Sociology of Education*, pages 328–348, 2002) present a similar measure, linking it to racially disparate tracking.

[10] Cook, Shadish and Wong's review ("Three conditions under which experiments and observational studies produce comparable

causal estimates: New findings from within-study comparisons," *Journal of Policy Analysis and Management*, 27(4):724–750, 2008) identified matching within local contexts as one attribute of quasi-experiments appearing to have successfully addressed selection bias. Bifulco's ("Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison" *Journal of Policy Analysis and Management*, 31(3):729–751, 2012) case study suggests that propensity matches from similar districts elsewhere within a state may be as good or better than same-district matches.

[11] In order that the variables used in matching precede the start of the program, at the time of matching the Evaluation Engine does not take into account whether a student's outcome test scores come from the standard or from alternate forms of the test. Instead, it attempts to avoid matching across boundaries of the following categories: not disabled; deaf and blind, and took alternate assessment in previous year; deaf and blind, but did not take alternate assessment in previous year (including no score recorded); autistic, and took alternate assessment in previous year; autistic, and did not take alternate assessment in previous year; other disabled, with likely mild or no cognitive impairment ("specific learning disability", "speech or language impairment", …); other disabled, with likely cognitive impairment ("traumatic brain injury", "intellectual disability", "multiple disability").

[12] The possibility of spillover effects is also addressed by the use of permutation methods for the significance tests reported with our overall average effect estimates; these remain statistically valid, in terms of their basic interpretations, in the presence of spillover, although spillover can limit their power.

[13] More formally, participation group students whose inclusive propensity score differs from each non-participant of similar disability status, to an extent that estimation error in the propensity score could not readily explain. This trimming of the sample is a by-product of broader requirements that matching be done within calipers of the inclusive PS (e.g., Rubin, D.B. and Thomas, N., "Matching using estimated propensity scores: Relating theory to practice," *Biometrics*, 52:249–64, 1996), the caliper width being a function of the magnitude of estimation error, and that matched counterparts not differ significantly from one another in terms of their estimated inclusive propensity scores, in multiplicity-corrected significance tests. Because the inclusive PS model coefficients have relatively large standard errors, participants are trimmed only if they are very different from the non-participant group.

[14] Gu, X.S. and Rosenbaum, P.R., "Comparison of multivariate matching methods: Structures, distances, and algorithms," *Journal of Computational and Graphical Statistics*, 2(4):405–420, 1993; Hansen, B.B. and Klopfer, S.O., "Optimal full matching and related designs via network flows," Journal of Computational and Graphical Statistics, 15(3):609–627, 2006.

[15] Hothorn T., Bretz F., and Westfall P., "Simultaneous inference in general parametric models," *Biometrical Journal*, 50(3):346–363, 2008.

[16] Specifically, the method of Abadie and Imbens ("A martingale representation for matching estimators," *Journal of the American Statistical Association*, 107(498):833–843, 2012), adapted to full matching (Rosenbaum, P.R., "A characterization of optimal designs for observational studies," *Journal of the Royal Statistical Society*, 53:597–610, 1991) and to covariance as well as variance estimation.

[17] In the tradition of Cornfield et al. ("Smoking and lung cancer: Recent evidence and a discussion of some questions," *Journal of the National Cancer Institute*, 22:173–203, 1959). The method adapts that of Hosman *et al*. ("The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder," Annals of Applied Statistics, 4(2):849–870, 2010) to the Evaluation Engine's estimation strategies, and to present sensitivity parameters in units comparable to those of Figure A1.

[18] The partial correlations in question can be recovered from a quantity of secondary interest in the planning of group randomized trials, Hedges and Hedberg's η_W ("Intraclass correlation values for planning group-randomized trials in education," *Educational Evaluation and Policy Analysis*, 29(1):60–87, 2007) . Hedges and Hedberg tabulate values of this quantity from nationally representative samples. While their results varied somewhat across grade and subject, the implied partial correlations generally fell in the neighborhoods of .75 (partial correlations with pretests) and .01 (demographic variables).

The Evaluation Engine does match on both demographic and prior achievement variables. Interestingly, however, the same reference values can be motivated by a study of effects of baseline variables that the Evaluation Engine does not address, namely classroom or teacher achievement data. Zhu, Jacob, Bloom and Xu ("Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information," *Educational Evaluation and Policy Analysis*, 34(1):45–68, 2012) quantitatively characterize the additional value of classroom-level data over and above individual- and school-level data, for the planning and analysis of randomized trials. In the lower grades, classroom-level data turns out to add relatively little from the trialist's perspective, while in high school it makes a meaningful contribution. Back-translating from their measures to ours, results again vary somewhat across grades, tests and data sources, but partial correlations of .1 typify the classroom data's additional contribution in the lower grades while .75 typifies their contribution to prediction of high school achievement outcomes.

4. USER DESCRIPTIONS

Name of intervention group

TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed

Description of intervention

Title of report

TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed

Description of analysis

Analysis Name: TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed
Study Name: TN Test 4 - G7 Y2011 Male Promotion 5pct Skewed
Intervention Start: 9/1/2010 12:00:00 AM
Intervention End: 1/1/1753 12:00:00 AM
Outcome Areas: readng_scr,readng_alt,engli2_tap,glmath_scr,glmath_alt,algbr1_tap,attend,status_inferred,persist_reported
Subgroups: lep_now
Grade Levels: 7

SUPPLEMENTAL STATISTICAL MATERIAL

Information about the match

STRUCTURE OF MATCHED SETS

Intervention students were matched to varying numbers of non-intervention students. A single intervention student could be matched to several control students (which we notate below as 1:x) or a single control student could be matched to several intervention students (notated x:1 below). It is never the case that multiple intervention students were simultaneously matched to multiple control students. The table below describes the number of matched sets for various configurations of treated and control students. There are diminishing returns to the amount of information added by including additional students in matched sets. This match configuration contained the equivalent of 2858.9 matched pairs.

WITHIN DISTRICT MATCHES

99.8
percent of all matches are within the same district.

.

COMBINED DIFFERENCES TEST OF BASELINE EQUIVALENCE

Goodness-of-fit tests for the statistical model that students matched to one another do not differ in terms of their underlying propensity scores (Hansen & Bowers, 2008, *Statist. Sci.* **23** (2) 219-236). Test pertaining to two propensity score specifications are presented, one pertaining to the main, individual level variables and the other to all individual level matching variables plus school demographics. (Variables contributing to each specification are presented next, under "Balance Tables.")

	CHI-SQUARE	DF	P-VALUE
main variables	70.14	23.00	0.00
inclusive	114.06	36.00	0.00

Balance Tables

These tables describe comparisons between the intervention study group and the matched controls. When calculating participant group means, each participant with a value (on the variable in question) who is matched to at least one non-participant with a value contributes in an equally weighted fashion. As this is being done, a simple average of available values among non-participants matched to that participant is calculated, generating a _matched comparison value_ on the variable in question for that participant. The tables compare participant means to means of these matched comparison values, in both cases taking unweighted means over non-missing values for which the matched counterpart is also not missing. Standard differences divide these differences by state cohort SDs of the variable; if matches have been made within multiple grades, then these standard differences are calculated separately within grades before being combined, with weights proportional to the size of the participant group in each grade, to give the number presented in the table. "Z statistics" scale these mean differences by the corresponding null standard error, revealing the contribution of a given variable to the chi-square statistics testing the hypothesis of perfect matching on the propensity score, above.

INCLUSIVE VARIABLES SPECIFICATION

	Participant group mean (not weighted)	Matched comparison mean	Standardized difference	*z* statistic
0				
(element weight)	1.00	1.00		
1				
Attendance rate, prior year	0.95	0.96	-0.06	-2.90
10				
Proficient, prior year: Proficient (or advanced)	0.44	0.44	0.01	0.47
11				
Score, prior year	746.02	746.18	-0.00	0.02
12				
School/subgroup scores, -1 yr	743.87	743.86	0.00	-0.02
13				
Promotion, prior year: Promoted	1.00	1.00	0.00	0.36
14				
Diversity index (excl. missing)	0.27	0.26	0.04	3.93
15				
% Asian	1.49	1.45	0.02	2.46
16				
% African American	23.75	23.58	0.01	0.60
17				
% Limited Eng. prof.	1.88	1.62	0.08	5.34
18				
% Hispanic/Latino	5.32	4.88	0.07	4.38
19				
% Special ed.	11.91	11.91	0.00	0.24
2				
Chronic absentee (attendance < .9): Yes	0.10	0.08	0.06	2.80
20				
% Eligible for FRL	44.89	44.27	0.03	2.20
21				
% Caucasian	68.94	69.59	-0.02	-2.68
22				
School-wide Title I: Yes	0.70	0.70	0.01	1.38
23				
Title I eligible: Yes	0.77	0.77	-0.00	0.39
39				
Gender: Female	0.00	0.00		
Gender: Male	1.00	1.00		
4				
Exam language, one year prior: English	0.98	0.97	0.02	0.98
40				
Limited-English proficient: Yes	0.01	0.01	0.04	2.40
41				
Race/ethnicity: American Indian or Alaska Native	0.00	0.00	0.01	0.53
Race/ethnicity: Asian	0.01	0.01	0.02	0.58
Race/ethnicity: Black or African American	0.24	0.24	-0.01	-0.46
Race/ethnicity: Hispanic/Latino	0.06	0.06	-0.02	-0.17
Race/ethnicity: Native Hawaiian or Other Pacific Islander	0.00	0.00	-0.02	-0.35
Race/ethnicity: White	0.69	0.68	0.01	0.31
42				

Special ed. (IDEA): Yes	0.16	0.16	0.00	1.00
5				
Proficient, prior year: Proficient (or advanced)	0.28	0.29	-0.00	-0.41
6				
Score, prior year	747.69	748.30	-0.02	-0.63
7				
School/subgroup scores, -1 yr	745.87	745.96	-0.00	-0.48
8				
Migrant, past 2 yrs: Yes	0.00	0.00	0.02	1.06
9				
Exam language, one year prior: English	0.97	0.97	0.01	0.10

MAIN VARIABLES ONLY

	Participant group mean (not weighted)	Matched comparison mean	Standardized difference	*z* statistic
0				
(element weight)	1.00	1.00		
1				
Gender: Female	0.00	0.00		
Gender: Male	1.00	1.00		
10				
Migrant, past 2 yrs: Yes	0.00	0.00	0.02	1.06
11				
Proficient, prior year: Proficient (or advanced)	0.44	0.44	0.01	0.47
12				
Score, prior year	746.02	746.18	-0.00	0.02
13				
Promotion, prior year: Promoted	1.00	1.00	0.00	0.36
14				
School-wide Title I: Yes	0.70	0.70	0.01	1.38
15				
Title I eligible: Yes	0.77	0.77	-0.00	0.39
2				
Limited-English proficient: Yes	0.01	0.01	0.04	2.40
3				
Race/ethnicity: American Indian or Alaska Native	0.00	0.00	0.01	0.53
Race/ethnicity: Asian	0.01	0.01	0.02	0.58
Race/ethnicity: Black or African American	0.24	0.24	-0.01	-0.46
Race/ethnicity: Hispanic/Latino	0.06	0.06	-0.02	-0.17
Race/ethnicity: Native Hawaiian or Other Pacific Islander	0.00	0.00	-0.02	-0.35
Race/ethnicity: White	0.69	0.68	0.01	0.31
4				
Special ed. (IDEA): Yes	0.16	0.16	0.00	1.00
5				
Attendance rate, prior year	0.95	0.96	-0.06	-2.90
6				
Chronic absentee (attendance < .9): Yes	0.10	0.08	0.06	2.80
8				
Proficient, prior year: Proficient (or advanced)	0.28	0.29	-0.00	-0.41
9				
Score, prior year	747.69	748.30	-0.02	-0.63

Additional Outcome Analysis Information

	PERMUTATION TEST STATISTIC	NULL SD OF TEST STATISTIC	DIFFERENCE OF MEANS (DEMOGRAPHICS ADJUSTED)	STANDARD DEVIATION
Grade-level reading, regular exam, scale score, current year (2011)	0.06	0.00	0.01	0.00
Grade-level math, regular exam, scale score, current year (2011)	-0.25	0.00	-0.02	0.00
Attendance rate, current year (2011)	-0.00	0.00	-0.00	0.00
Promotion/retention status, as inferred, current year (2011)	0.00	0.00	0.00	0.00
Persistence status, as reported, current year (2011)	0.00	0.00	0.00	0.00
Grade-level reading, regular exam, scale score, one year later (2012)	-0.39	0.00	-0.01	0.00
Grade-level math, regular exam, scale score, one year later (2012)	-0.52	0.00	-0.01	0.00
Attendance rate, one year later (2012)	-0.00	0.00	-0.01	0.00
Number of years promoted to a higher grade, as inferred, one year later (2012)	0.01	0.00	0.00	0.00
Persistence status, as reported, one year later (2012)	-0.00	0.00	-0.00	0.00
English 2 (EOC), exam taken and passed by two years later (2013)	0.00	0.00	0.00	0.00
Algebra 1 (EOC), exam taken and passed by two years later (2013)	0.03	0.00	0.03	0.00
Attendance rate, two years later (2013)	0.00	0.00	0.02	0.00
Number of years promoted to a higher grade, as inferred, two years later (2013)	0.02	0.00	0.01	0.00
Persistence status, as reported, two years later (2013)	-0.00	0.00	-0.00	0.00

PER-GRADE AVERAGES

VARIABLE	GRADE	MEAN		SD		PROFICIENCY		ATTRITION	
		TREATED	CONTROL	TREATED	CONTROL	TREATED	CONTROL	TREATED	CONTROL
Grade-level reading, regular exam, scale score, current year (2011)	7	750.549	750.460	37.171	26.663	0.427	0.415	0.125	0.172
Grade-level math, regular exam, scale score, current year (2011)	7	755.294	756.159	42.775	29.775	0.338	0.356	0.125	0.173
Attendance rate, current year (2011)	7	0.951	0.951	0.050	0.033	NA	NA	0.001	0.001
Promotion/retention status, as inferred, current year (2011)	7	0.992	0.987	0.089	0.066	NA	NA	0.032	0.067
Persistence status, as reported, current year (2011)	7	1.000	1.000	0.000	0.011	NA	NA	0.000	0.000
Grade-level reading, regular exam, scale score, one year later (2012)	7	754.203	754.558	36.867	24.855	0.449	0.447	0.189	0.270
Grade-level math, regular exam, scale score, one year later (2012)	7	762.150	762.784	36.638	27.003	0.379	0.391	0.289	0.414
Attendance rate, one year later (2012)	7	0.951	0.951	0.056	0.033	NA	NA	0.032	0.066
Number of years promoted to a higher grade, as inferred, one year later (2012)	7	1.981	1.976	0.136	0.091	NA	NA	0.074	0.144
Persistence status, as reported, one year later (2012)	7	0.999	1.000	0.034	0.008	NA	NA	0.032	0.066
English 2 (EOC), exam taken and passed by two years later (2013)	7	0.006	0.002	0.077	0.035	NA	NA	0.087	0.168
Algebra 1 (EOC), exam taken and passed by two years later (2013)	7	0.038	0.004	0.236	0.037	NA	NA	0.884	0.932
Attendance rate, two years later (2013)	7	0.949	0.947	0.063	0.038	NA	NA	0.068	0.132
Number of years promoted to a higher grade, as inferred, two years later (2013)	7	1.961	1.946	0.194	0.150	NA	NA	0.112	0.204
Persistence status, as reported, two years later (2013)	7	0.996	1.000	0.059	0.012	NA	NA	0.068	0.132

ORDINAL OUTCOMES

These tables display the percentage of treated and control units in each of the category conditions. Per-grade, the treated values should sum to 100%. Separately, the control values sum to 100%.

Promotion/retention status, as inferred, current year (2011)

RETAINED		PROMOTED	
TREATMENT	CONTROL	TREATMENT	CONTROL
0.8	1.25	99.2	98.75

Number of years promoted to a higher grade, as inferred, one year later (2012)

RETAINED BOTH YEARS		PROMOTED ONE (OF TWO) YEARS		PROMOTED BOTH YEARS	
TREATMENT	CONTROL	TREATMENT	CONTROL	TREATMENT	CONTROL
0	0.05	1.9	2.33	98.1	97.62

English 2 (EOC), exam taken and passed by two years later (2013)

DID NOT TAKE EXAM		TOOK BUT DID NOT PASS EXAM		PASSED EXAM	
TREATMENT	CONTROL	TREATMENT	CONTROL	TREATMENT	CONTROL
99.4	99.85	0.6	0.1	0	0.05

Algebra 1 (EOC), exam taken and passed by two years later (2013)

DID NOT TAKE EXAM		TOOK BUT DID NOT PASS EXAM		PASSED EXAM	
TREATMENT	CONTROL	TREATMENT	CONTROL	TREATMENT	CONTROL
97.16	99.57	1.9	0.43	0.95	0

Number of years promoted to a higher grade, as inferred, two years later (2013)

RETAINED ALL THREE YEARS		PROMOTED ONE (OF THREE) YEARS		PROMOTED TWO (OF THREE) YEARS	
TREATMENT	CONTROL	TREATMENT	CONTROL	TREATMENT	CONTROL
0	0.11	3.9	5.21	96.1	94.68

Logs

During the course of the evaluation, the statistical system will log various information explaining decisions made during matching.

NAME	TIME	MESSAGE
user	Mon Apr 17 20:43:10 2017	We do not currently have data to provide analysis on readng_alt for 2011.
user	Mon Apr 17 20:43:10 2017	We do not currently have data to provide analysis on glmath_alt for 2011.
user	Mon Apr 17 20:43:10 2017	We do not currently have data to provide analysis on readng_alt for 2012.
user	Mon Apr 17 20:43:10 2017	We do not currently have data to provide analysis on glmath_alt for 2012.
user	Mon Apr 17 20:43:11 2017	We do not currently have data to provide analysis on readng_alt for 2013.
user	Mon Apr 17 20:43:11 2017	We do not currently have data to provide analysis on glmath_alt for 2013.
user	Mon Apr 17 20:43:11 2017	Starting a job for TN for academic year 2010-09-01
user	Mon Apr 17 20:43:11 2017	User has requested 1825 intervention subjects for this job.
user	Mon Apr 17 20:43:11 2017	User limited job to grades 7
user	Mon Apr 17 20:43:12 2017	Found 1820 intervention subjects in the requested grade level (7)
user	Mon Apr 17 20:43:14 2017	In deciding which schools to use, using variables: Diversity index (excludes missing), Percent of Asian students, Percent of black or African American students, Percent of Limited-English proficient students, Percent of Hispanic or Latino students, Percent of special education students, Percent of students eligible for free or reduced-price lunch, Percent of white students, School-wide Title I, Title I Eligible School, Grade-level math exam, average 7th grade smoothed score, current year, Grade-level reading exam, average 7th grade smoothed score, current year
user	Mon Apr 17 20:43:25 2017	Calipered at the school level on diversity_sch, pctasian_sch, pctblack_sch, pctell_sch, pcthisp_sch, pctspeced_sch, pcttotfrl_sch, pctwhite_sch, glmath_smhsch_gl7_p0, readng_smhsch_gl7_p0 with a width of 3 standard devations.
user	Mon Apr 17 22:50:16 2017	Matching completed, took approx. 2 hour(s) and 7 minute(s).

JobGUID: 1c605bdc-815c-4a6b-8b32-99fb0233f043
Git Status: