Napa County Office of Education
CalSERVES AmeriCorps Impact Evaluation 2020

Prepared by JBS International, Inc.
October 26, 2020

Introduction

The Napa County Office of Education's CalSERVES AmeriCorps Program (CalSERVES) offers individualized tutoring and mentoring to approximately 260–450 students each semester in after school programs and school club activities for the period covered in this analysis. The Napa County Office of Education contracted with JBS International, Inc. to conduct a quasi-experimental evaluation of the CalSERVES program using propensity scores to match program participants with students who did not participate in the program but were from the same schools.[1] This evaluation uses data from the STAR360 literacy (STAR Early Literacy and STAR Reading) and STAR360 math assessments for the 2018-2019 school year, and fall semester of the 2019–2020 school year and the Strengths and Difficulties Questionnaire (SDQ) from 2018–2019 school year. Assessment scores and school and demographic information were included in analyses with a sample size of approximately 100–200 program participants and 1,400–1,500 non-participants per semester for literacy, 40–60 program participants and 250–600 non-participants per semester for math, and 57 program participants and 409 non-participants for the SDQ which was used to measure mentoring outcomes for students participating during the2018–2019 school year. The study assessed the following evaluation questions:

1. Do CalSERVES tutoring recipients demonstrate improved academic performance as measured by STAR360 literacy and math assessments when compared with a propensity score matched group of students who do not receive tutoring services?

2. Do CalSERVES mentoring recipients demonstrate improved social and behavioral skills when compared with a propensity score matched group of students who do not receive mentoring services?

Methods

The datasets provided to JBS by the Napa County Office of Education included assessment data for STAR360 literacy, DIBELS, RI, MI, and STAR360 math. Due to the small sample sizes of students who received tutoring and took the DIBELS, RI, and MI assessments, these scores are not part of the analysis. There were no students in Fall 2019 that received tutoring and took the MI assessment. Table 1 shows the number of students with complete data who received tutoring.

Data Sources Used for Analysis. We used three primary datasets for the analysis:

[1] Please note that tutoring is assigned at the student level, with both treatment and comparison students available within the same school.

1. **Literacy**. This dataset contained the STAR360 literacy (STAR Early Literacy and STAR Reading) assessment pre- and post-test percentile and performance level for the 2018–2019 school year and Fall 2019, CalSERVES literacy tutoring participation data, and demographic data[2] for students at six schools in the Napa Valley Unified School District (four schools did not have data for STAR360 literacy).[3]

2. **Math**. This dataset contained the STAR360 math assessment percentile and performance level pre- and post-test percentile and performance level, CalSERVES STEM tutoring participation data, and demographic data[4] for students at two schools in the Napa Valley Unified School District (four schools did not have data for STAR360 math).[5]

3. **SDQ**. This dataset contained SDQ total scores and subscale scores for the 2018–2019 school year, CalSERVES mentoring participation data, and demographic data for students at three middle schools.[6]

Propensity Score Matching. We used propensity scores to create matched comparison groups for CalSERVES students in each of the three datasets above. Propensity score matching is a quasi-experimental design that matches students in a treatment group (in this case those receiving at least one CalSERVES literacy, STEM, or mentoring session) to students in a comparison group based on demographic characteristics and characteristics likely to predict program participation (such as prior literacy, math performance, or prior SDQ score) to balance the two groups of students. In cases where random assignment is not feasible, PSM is a well-validated alternative that can generate a group of treatment and comparison students with a statistically equalized likelihood of program participation. The matched sample is based on measured variables; however, there are potential other factors which might affect the likelihood of participating in the CalSERVES tutoring program and that affect literacy and math skills which are not measured, and therefore could not be accounted in the analysis.

It is important to note that the comparison students were all getting service within the school day from (at a minimum) instructional assistants and literacy paraprofessionals. Thus, the comparisons made are not between AmeriCorps tutored students and unserved students, but between students served by AmeriCorps and students provided additional support from school supervised assistants and paraprofessionals. We conducted the propensity score matching (PSM) for literacy and math for each semester Fall 2018, Spring 2019, and Fall 2019 separately. We conducted the PSM for the SDQ data for the entire 2018–2019 school year.

[2] Demographic data included ethnicity, race, English-language learner status, and special education status.
[3] The following schools provided STAR360 literacy data: Bellevue Elementary, Kawana Springs Elementary, Meadow View Elementary, RL Stevens Elementary, Taylor Mountain Elementary, and Wright Charter. Napa Valley Language Academy, Phillips Elementary, Shearer Elementary, and Snow Elementary did not provide data for STAR360 literacy. Grade level information was also provided for each student.
[4] Demographic data included ethnicity, English-language learner status, and special education status.
[5] The following schools provided STAR360 math data: RL Stevens Elementary and Wright Charter (Fall 2018 and Spring 2019 data contains only Wright Charter). Napa Valley Language Academy, Phillips Elementary, Shearer Elementary, and Snow Elementary did not provide data for STAR360 math.
[6] Schools with SDQ data included: Harvest Middle School, Redwood Middle School, and Silverado Middle School.

After successful matching, any differences observed between treatment and comparison students can generally be attributed to program participation.[7]

Students were matched one-to-one on the following variables using a nearest neighbor matching algorithm with a .25 standard deviation caliper:

- STAR360 literacy and math pre-test percentile (for the literacy and math datasets respectively)
- SDQ total score (for the SDQ dataset)
- Gender (only available in the SDQ dataset)
- Race/ethnicity (White non-Hispanic, Hispanic, or Other non-Hispanic)
- English-language learner (ELL) Status
- Special education status
- School attended

It was not possible to do an exact matching on grade due to the small sample number of students who received tutoring within each grade. Table 2 shows the number of students who had tutoring and complete data by grade.

Literacy Tutoring PSM Results. The STAR360 literacy dataset for Fall 2018 included 109 students in the literacy tutoring program and 1,488 comparison students that did not participate in the tutoring program. In this dataset, 108 tutoring students and 1,459 comparison students had complete data and were eligible for the propensity score matching (leaving 99% of treatment students and 98% of comparison students for matching for Fall 2018 semester). The PSM process yielded matches for 106 tutoring students and 106 comparison students (i.e., one 'matched' comparison student was identified for each program participant).

The literacy dataset for Spring 2019 included 211 students in the tutoring program and 1,467 comparison students that did not participate in the tutoring program. In this dataset, 204 tutoring students and 1,405 comparison students had complete data and were eligible for propensity score matching (leaving 97% of treatment students and 96% of comparison students for matching for Spring 2019 semester). The PSM process yielded matches for 204 program students and 204 comparison students.

The literacy dataset for Fall 2019 included 226 students in the tutoring program and 1,543 comparison students that did not participate in the tutoring program. In this dataset, 223 tutoring students and 1,537 comparison students had complete data and were eligible for propensity score matching (leaving 99% of treatment students and 99.6% of comparison students for matching for Fall 2019 semester). The PSM process yielded matches for 223 program students and 223 comparison students.

Prior to matching, students that participated in the literacy tutoring program had significantly lower percentile than their non-tutored peers. Literacy tutoring students were also more likely to

[7] This assumes that the match was successful in eliminating pre-test differences between the treatment and comparison students and there were no significant unobserved variables excluded from the match. See Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, *46*(3), 399-424.

be Hispanic in Spring and Fall 2019, more likely to be English-language learners in all three semesters, and more likely to be of other ethnicity in Fall 2019 (Tables 4–6). These comparisons indicate that CalSERVES is successfully engaging a high-need population in their literacy tutoring program. After the matching process, all significant differences between students in the tutoring program and the comparison students were eliminated (Tables 4–6).

Math Tutoring PSM Results. The STAR360 math dataset for Fall 2018 included 50 students in the STEM tutoring program and 262 comparison students that did not participate in the STEM tutoring program. In this dataset, 47 STEM tutoring students and 256 comparison students had complete data and were eligible for propensity score matching (leaving 94% of treatment students and 98% of comparison students for matching for Fall 2018). The propensity score matching process yielded matches for 46 STEM program students and 46 comparison students (i.e., one 'matched' comparison students was identified for each program participant).

The math dataset for Spring 2019 included 47 students in the STEM tutoring program and 267 comparison students that did not participate in the STEM tutoring program. In this dataset, 43 STEM tutoring students and 252 comparison students had complete data and were eligible for propensity score matching (leaving 91% of treatment students and 94% of comparison students for matching for Spring 2019). The PSM process yielded matches for 43 STEM program students and 43 comparison students.

The math dataset for Fall 2019 included 62 students in the STEM tutoring program participants and 608 comparison students that did not participate in the STEM tutoring program. In this dataset, 62 STEM tutoring students and 606 comparison students had complete data and were eligible for propensity score matching (leaving 100% of treatment students and 99.7% of comparison students for matching for Fall 2019). The PSM process yielded matches for 62 STEM program students and 62 comparison students.

Prior to matching, students that participated in the STEM tutoring program had significantly lower math percentile than their non-tutored peers. STEM tutoring students were also more likely to be Hispanic and an English-language learner in Fall 2018 and Spring 2019 (See Tables 7–9). These comparisons indicate CalSERVES is successfully engaging a high-need population with their STEM tutoring program. After the matching process, all significant differences between the program group and the comparison group were eliminated. One exception was in Fall 2018, where the percentage of special education students were significantly higher in the comparison group, but that should be interpreted with caution, as there was only one student with special education status in the matched sample.

Mentoring (SDQ) PSM Results. The SDQ dataset included 61 students in the mentoring program and 422 comparison students that did not participate in the mentoring program. Out of this dataset, 57 mentoring students and 409 comparison students had complete data and were eligible for propensity score matching (leaving 93% of treatment students and 97% of comparison students for matching). The PSM process yielded matches for 57 mentoring students and 57 comparison students (i.e., one 'matched' comparison student was identified for each mentoring student).

Prior to matching, students who participated in the mentoring program were more likely to be of 'Other' ethnicity, fewer females, and more students with special education status (Table 10). After the PSM process, all significant differences between the program students and the comparison group students were eliminated.

Impact Analyses. We used multivariate linear and logistic regression approaches using the treatment and matched comparison group to answer the evaluation questions (i.e., do tutored or mentored students show improvement versus comparison students) for Fall 2018, Spring 2019 and Fall 2019 separately. To test for this improvement, we used two outcome variables:

1. **Performance Level (PL) Improvement (for literacy and math only)**. This dichotomous variable showed whether the students' overall performance level (assessed as being far below standard [less than 25th percentile], below standard [25–29th percentile], meets standard [50–74th percentile], or exceeds standard [75–99th percentile]) improved from pre- to post-test within each semester for the literacy or math test. For example, we coded students as improved if they moved from "far below standard" to "below standard"; we coded students as did not improved if their performance levels stayed the same or declined.[8] We used multivariate logistic regressions to predict PL improvement (yes/no) while controlling for grade, race/ethnicity, ELL status, special education status, and school. We conducted separate analyses for Fall 2018, Spring 2019, and Fall 2019.

2. **Post-Test Percentile Scores (literacy and math) and Raw Scores (SDQ)**. The percentile scores for literacy and math, and the raw score for SDQ are continuous variables. When controlling for pre-test scores, positive changes in this outcome can be interpreted as improvements over time. We used multivariate linear regressions to predict post-test scores while controlling for pre-test scores, gender (SDQ only), ethnicity, ELL status, special education status, and school.

   For the math sample, we did not include special education status and school in the logistic regression and linear regression models due to the few number of students in the treatment group; in this case the regression models could not be estimated. Specifically, in Fall 2018, there was only one student with special education status. In Fall 2018 and Spring 2019, Wright Charter was the only school with STAR360 math data. To obtain comparable results across the three semesters, we dropped special education and school from all three models.

Literacy Tutoring Results

PL Improvement and Literacy Percentile Results. The pre- and post-test mean percentile are shown in Table 3. In Fall 2018, students that received any literacy tutoring (one or more sessions) did not show any statistically significant differences from those who did not in the likelihood of literacy performance improvements (Table 11). They had significantly lower percentile at post-test (B = -4.78, SE = 2.00, t = -2.40, p = .02; Table 12). In Spring and Fall

[8] Improvement in performance levels was selected as an appropriate outcome since 96–99% of students in the final matched literacy sample and 91–96% in the matched math sample had room for performance level improvement (i.e., far below or below standard).

2019, students that received any tutoring did not show any statistically significant differences from those who did not in the likelihood of literacy performance improvements or percentile at post-test (Tables 11 and 12).

STEM Tutoring Math Results

PL Improvement and Math Percentile Results. The pre- and post-test mean percentile are shown in Table 3. In Fall 2018, students that received any STEM tutoring (one or more sessions) did not show any statistically significant differences from those who did not in the likelihood of math performance improvements (Table 13). They had significantly lower percentile than those who did not at post-test (B = -12.64, SE = 3.90, t = -3.24, p = .002; Table 14). In Spring and Fall 2019, students that received any STEM tutoring did not show any statistically significant differences from those who did not in the likelihood of math performance improvements or percentile at post-test (Tables 13 and 14).

Mentoring SDQ Results

SDQ Results. The pre- and post-test mean total score are shown in Table 3. Students that received any mentoring (one or more sessions) did not show any statistically significant differences in their total SDQ scores at post-test (see Table 15). Additionally, there were no statistically significant differences in post-test SDQ scores or the two subscales of the SDQ (difficulties and prosocial behaviors).

Conclusions and Next Steps

Overall, the results indicate that CalSERVES is successfully engaging a high-need population in its services, and students participating in CalSERVES tutoring show similar growth to other high need students receiving school-provided support services. In fact, the comparison group in the unmatched sample scored higher in the pretest in all data sets, indicating that the school is targeting the lowest performing students for the CalSERVES program. Superintendents, principals, and teachers share their positive perceptions of the program in preparing students for additional intervention services, which may play an important role in the findings of this study.

Literacy Tutoring Results. The CalSERVES literacy tutoring analysis results suggest students in the program were no different when compared to a matched comparison group of students who did not participate in that program but received school interventions. This program provides an important service to the school community as it supports students who otherwise would not be served due to resource constraints. Similar to the comparison group, students participating in the program showed percentile growth while not meeting the threshold for performance level improvement (Fall 2018). Based on an understanding of the program, the findings may have two possible explanations: (1) there may be opportunities to adjust tutoring services' content or activities to better serve this population; (2) similar students who are not receiving CalSERVES services (i.e., comparison students) are receiving school day services which are not measured in the data used for the analysis. In other words, there are no 'no treatment' comparison students since all the students who need help get help from their school.

This explanation is supported by the fact that CalSERVES participants are students who are flagged by the school but who do not receive services due to limited resources. At the same time, comparison students were also flagged by the school but are receiving additional in-school tutoring and support. Students who receive more intensive in-school tutoring and support may be more likely to show improvements than students who only receive the tutoring and support provided by CalSERVES. To investigate this, future evaluations might obtain more detailed information on the services that students (both treatment and comparison students) are obtaining during the school day to ensure that comparison students are not contaminated by any outside services, or, alternately to confirm that students served primarily by CalSERVES and those served primarily by in-school professional and para-professional staff are performing as similarly as the current data shows. Another explanation is that the effects could not be detected with the small sample sizes available for the analysis. If this is the case, further analysis with larger sample sizes is needed to detect small effects (if any). The time frame is fairly short (just a semester); perhaps the students build a deeper literacy foundation which will sustain them on a continuous growth path in later grades – something which could be tested with longitudinal data (multiple years of data on the same students) as the students participating in CalSERVES advance to later grades.

Math Tutoring Results. Like the literacy tutoring results, students in the program did not perform differently than comparison students who received school services.. Similar to the literacy tutoring results, there were mixed results in Fall 2018. Students who received tutoring in Fall 2018 showed percentile growth but did not show statistically significant differences in performance level improvement. As with the literacy results, the math sample also had small sample sizes, and thus, may not be sensitive enough to detect small effects. Future analysis with a large sample size is needed to detect these effects. Furthermore, as noted above, there may not be a no 'no treatment' comparison students. Therefore, it would also be helpful to assess what alternative services are being offered to comparison students so that future evaluations can rule out alternative service receipt as a potential limitation for results.

Mentoring SDQ Results. Overall, there were no statistically significant differences between the SDQ post-test scores of students who participated in CalSERVES mentoring and those who did not. Students were identified for additional support and mentoring by teachers, principals and counselors. Other students who scored similarly may have been served through in-school services and wellness centers. There were also no statistically significant differences in the two subscale scores, difficulties and prosocial behaviors. These findings suggest that students who participate in the mentoring program do not appear to differ in their SDQ performance compared to those who participate in other school interventions in a matched comparison sample. Additional research would be useful to explore what other services non-mentoring students are receiving and to determine if mentoring is having an impact in areas not measured by the SDQ. For example, while mentored students may not be more likely to show improvements on the SDQ, it is possible they may be experiencing other positive benefits that were not measured in the SDQ. The SDQ only measured emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behavior.

Limitations and Considerations

There are a number of limitations to take into consideration when looking at the results of this evaluation. First, the fact that it was not possible to compare service recipients to a true no treatment group is a consequential limitation of this evaluation. All students who are in need of services are either served by school specialists or by AmeriCorps tutors, so some level of improvement would be expected for all.

A second limitation is PSM and the amount of information available in the data to conduct the matching. The PSM design is well-validated; however, the matching process assumes there are no unobserved differences between the two groups conditional on the observed variables. The matching procedure is limited because the data available for the analysis potentially does not include all relevant measures required for this procedure. For instance, as noted in the results, there are no measures of additional services that non-tutoring students receive during the school day, although it is anticipated that all comparison students below the 35th percentile at pretest receive some services. The data also do not include measures on learning resources available at home to support literacy and math skills. Since additional services, and in particular learning resources, available at home are important contributors to learning outcomes, the lack of these measures suggests that the two groups could still differ on those unobserved but important measures. This would then contribute to differences in learning outcomes that are not captured in the model. As such, it is important for the matching procedure to include all variables related to the likelihood of being assigned to CalSERVES and literacy and math scores. Given these data limitations, it may be useful to consider either obtaining additional data on treatment and comparison students, using an alternative approach, or using supplementary approaches moving forward.

A third limitation is the use of single measures to assess program impacts. Further, using grade-based math and literacy assessments to show performance level improvements for below grade-level students is a limitation of the single measure used for tutoring. Tutors' work with students focuses on their areas of need, which may be below grade level, such as phonics for a fourth grader. As a result, the assessment is often unlikely to capture the actual progress made by students as a result of tutoring if it's not a skill assessed at that grade level. Additionally, districts using different types of assessments limited the sample sizes, including the loss of one district which did not use the STAR360 assessment. Furthermore, four schools did not have data for STAR360 literacy, four schools did not have data for STAR360 math, and only three schools had mentoring.

A fourth limitation is limited sample size, which limited our ability to detect small effects. The best assessment with the largest sample size we had was STAR360 literacy and math. There were 108, 204, and 223 students who had literacy tutoring (and took the STAR360 literacy assessment) in the fall 2018, spring 2019, and fall 2019 semesters; 47, 43, 62 students who had math tutoring in the fall 2018, spring 2019, and fall 2019 semesters; and 57 students who had mentoring in the 2018–2019 school year. The limited treatment sample size also limited the number of potential matches in PSM.

Finally, another crucial limitation is the fact that schools want the program to serve as many students as possible, which has resulted in semester-long tutoring versus whole year or even multiple years of tutoring. Given the high need population that the program serves, limiting

tutoring to one semester may not be enough support for many students to show performance level improvements on district, grade-based assessments.

Recommendations

The descriptive statistics of pre- and post-test scores by number of sessions show that for math, on average, the more tutoring sessions students had the larger their percentile increase at post-test. Future studies can further explore dosage effects, such as more granular levels of sessions (e.g., 1–10, 11–20, 21–30, etc.) and testing the different levels of sessions as covariates in logistic and linear regression models. Anecdotal evidence indicates that this program prepares students for school success, establishes positive social-emotional relationships with other students and adults, and provides necessary academic support.

Based on the findings from this study that participating students do not perform significantly better or worse than non-participating students who are served by paraprofessional and other staff during the school day, it will be important to identify the exact services that participating and non-participating students are receiving and to compare the quality, intensity and corresponding costs of those services. Future evaluation plans should include a formative component to identify exactly how students are identified for the tutoring program vs. in-school service, the extent to which students are receiving the intended type, quality and amount of tutoring, and a cost comparison component to identify the relative costs of AmeriCorps tutoring vs. the in-school services provided to students who are similar at pre-test.

Given the limited data available on both treatment and comparison students, and the extent to which the treatment students are performing below grade level at pre-test, it may also be useful to explore the use of a time series analysis where each student serves as their own comparison case, using standardized assessments matched to their educational level and area of focus. This would require the identification and measurement of students in the semester before participation in the tutoring program, and measurement of students during the semester following tutoring.

Table 1

Number of Students with Complete Data and Received Tutoring or Mentoring, by Assessment and Semester

	Fall 2018	Spring 2019	Fall 2019	2018–2019 School Year
STAR360 Literacy	108	204	223	
DIBELS	20	52	69	
RI	31	41	54	
STAR360 Math	47	43	62	
MI	40	57	0	
SDQ	--	--	--	57

Note: STAR360 Literacy includes STAR Early Literacy and STAR Reading. SDQ dataset was for the entire school year.

Table 2

Number of Students with Complete Data and Received Tutoring or Mentoring, by Grade

Grade	STAR360 Literacy			STAR360 Math			SDQ
	Fall 2018	Spring 2019	Fall 2019	Fall 2018	Spring 2019	Fall 2019	2018–2019 School Year
1	7	26	25	--	--	--	--
2	13	33	29	2	2	4	--
3	46	64	57	4	4	11	--
4	15	32	52	4	3	10	--
5	18	26	44	15	14	10	--
6	9	23	16	11	11	20	22
7	--	--	--	6	5	3	16
8	--	--	--	5	4	4	19
Total	108	204	223	47	43	62	57

Note: SDQ dataset was for the entire school year.

Table 3

Mean Pretest and Posttest Scores of Matched Students for Literacy and Math Tutoring Program Participants, by Semester

	Treatment			Comparison		
	Pretest Mean	Posttest Mean	Change Mean	Pretest Mean	Posttest Mean	Change Mean
STAR360 Literacy						
Fall 2018	9.29	16.22	6.93	10.10	20.66	10.56
Spring 2019	16.70	18.37	1.67	18.67	20.56	1.89
Fall 2019	11.70	18.26	6.56	11.52	20.60	9.08
STAR360 Math						
Fall 2018	19.48	22.50	3.02	19.87	36.35	16.48
Spring 2019	23.17	22.13	-1.04	23.77	27.64	3.87
Fall 2019	24.69	38.32	13.63	25.68	38.74	13.06
SDQ						
2018–2019 School Year	19.98	19.57	-0.41	19.84	19.96	0.12

Note: STAR360 literacy and math are percentile scores. SDQ is total score. Change is posttest minus pretest.

Table 4

Summary of Propensity Score Matching Results for Literacy Tutoring Program Participants (Fall 2018)

	UNMATCHED Sample					1:1 MATCHED Sample				
	Treatment (N=108)		Comparison (N=1459)		Significant Difference?[9]	Treatment (N=106)		Comparison (N=106)		Significant Difference?
Variable	Mean/ Percent	SE	Mean/ Percent	SE		Mean/ Percent	SE	Mean/ Percent	SE	
Percentile (Pretest)	9.14	0.92	32.34	0.65	***	9.29	0.89	10.10	0.99	n.s.
Hispanic	87%	3%	80%	1%	n.s.	89%	3%	89%	3%	n.s.
Other Ethnicity[10]	6%	2%	10%	1%	n.s.	5%	2%	7%	2%	n.s.
English Learner	82%	4%	54%	1%	***	82%	4%	83%	3%	n.s.
Special Ed Student	4%	2%	5%	1%	n.s.	4%	2%	3%	2%	n.s.
Bellevue ES	26%	4%	15%	<1%	**	26%	3%	27%	3%	n.s.
Meadow ES	9%	3%	20%	<1%	**	9%	2%	10%	2%	n.s.
RL Stevens	11%	3%	16%	<1%	n.s.	11%	2%	13%	2%	n.s.
Taylor Mountain ES	14%	3%	21%	<1%	n.s.	14%	2%	11%	2%	n.s.
Wright Charter[11]	10%	3%	19%	<1%	*	10%	2%	11%	2%	n.s.

*$p < .05$. **$p < .01$. ***$p < .001$.

Note: SE = Standard error of mean/percent.

The sample includes only students who took the STAR360 literacy assessment (the most common literacy assessment in the data). Students who took the RI (N = 31 treatment cases with complete data) and DIBELS (N = 20 treatment cases with complete data) assessments were excluded.

[9] Tests for statistical significance were conducted using regressions that mirror the conclusions of a two-tailed independent samples t-tests for each variable.

[10] White race/ethnicity was not included in the matching since it would be redundant with the Other ethnicity variable, but it served as the reference category.

[11] Kawana ES was not included in the matching since it would be redundant with the other school variables, but it served as the reference category.

Table 5

Summary of Propensity Score Matching Results for Literacy Tutoring Program Participants (Spring 2019)

	UNMATCHED Sample					1:1 MATCHED Sample				
	Treatment (N=204)		Comparison (N=1405)		Significant Difference? [12]	Treatment (N=204)		Comparison (N=204)		Significant Difference?
Variable	Mean/ Percent	SE	Mean/ Percent	SE		Mean/ Percent	SE	Mean/ Percent	SE	
Percentile (Pretest)	16.70	1.09	37.62	0.68	***	16.70	1.07	18.67	1.18	n.s.
Hispanic	88%	2%	80%	1%	**	88%	2%	89%	2%	n.s.
Other Ethnicity[13]	5%	2%	9%	1%	n.s.	5%	2%	4%	1%	n.s.
English Learner	82%	3%	55%	1%	***	82%	3%	84%	3%	n.s.
Special Ed Student	2%	1%	5%	1%	n.s.	2%	1%	1%	1%	n.s.
Bellevue ES	31%	3%	14%	<1%	***	31%	2%	37%	2%	n.s.
Meadow ES	11%	2%	20%	<1%	**	11%	2%	11%	2%	n.s.
RL Stevens	9%	2%	14%	<1%	*	9%	1%	6%	1%	n.s.
Taylor Mountain ES	18%	3%	22%	<1%	n.s.	18%	2%	19%	2%	n.s.
Wright Charter[14]	5%	2%	20%	<1%	***	5%	1%	5%	1%	n.s.

*$p < .05$. **$p < .01$. ***$p < .001$.

Note: SE = Standard error of mean/percent. n.s. = Nonsignificant.

The sample includes only students who took the STAR360 literacy assessment (the most common literacy assessment in the data). Students who took the RI (N = 41 treatment cases with complete data) and DIBELS (N = 52 treatment cases with complete data) assessments were excluded.

[12] Tests for statistical significance were conducted using regressions that mirror the conclusions of a two-tailed independent samples t-tests for each variable.

[13] White race/ethnicity was not included in the matching since it would be redundant with the Other ethnicity variable, but it served as the reference category.

[14] Kawana ES was not included in the matching since it would be redundant with the other school variables, but it served as the reference category.

Table 6

Summary of Propensity Score Matching Results for Literacy Tutoring Program Participants (Fall 2019)

	UNMATCHED Sample					1:1 MATCHED Sample				
	Treatment (N=223)		Comparison (N=1537)		Significant Difference? [15]	Treatment (N=223)		Comparison (N=223)		Significant Difference?
Variable	Mean/ Percent	SE	Mean/ Percent	SE		Mean/ Percent	SE	Mean/ Percent	SE	
Percentile (Pretest)	11.70	0.81	33.24	0.65	***	11.70	0.80	11.52	0.85	n.s.
Hispanic	91%	2%	79%	1%	***	91%	2%	88%	2%	n.s.
Other Ethnicity[16]	5%	2%	10%	1%	*	5%	1%	7%	2%	n.s.
English Learner	76%	3%	44%	1%	***	76%	3%	75%	3%	n.s.
Special Ed Student	84%	2%	67%	1%	***	84%	2%	85%	2%	n.s.
Bellevue ES	11%	2%	16%	<1%	n.s.	11%	2%	12%	2%	n.s.
Meadow ES	20%	2%	17%	<1%	n.s.	20%	2%	19%	2%	n.s.
RL Stevens	12%	2%	19%	<1%	**	12%	2%	11%	2%	n.s.
Taylor Mountain ES	20%	2%	19%	<1%	n.s.	20%	2%	20%	2%	n.s.
Wright Charter[17]	7%	2%	21%	<1%	***	7%	1%	19%	1%	***

*$p < .05$. **$p < .01$. ***$p < .001$.

Note: SE = Standard error of mean/percent. n.s. = Nonsignificant.

The sample includes only students who took the STAR360 literacy assessment (the most common literacy assessment in the data). Students who took the RI (N = 54 treatment cases with complete data) and DIBELS (N = 69 treatment cases with complete data) assessments were excluded.

[15] Tests for statistical significance were conducted using regressions that mirror the conclusions of a two-tailed independent samples t-tests for each variable.

[16] White race/ethnicity was not included in the matching since it would be redundant with the Other ethnicity variable, but it served as the reference category.

[17] Kawana ES was not included in the matching since it would be redundant with the other school variables, but it served as the reference category.

Table 7

Summary of Propensity Score Matching Results for Mathematics Tutoring Program Participants (Fall 2018)

	UNMATCHED Sample					1:1 MATCHED Sample				
	Treatment (N=47)		Comparison (N=256)		Significant Difference? [18]	Treatment (N=46)		Comparison (N=46)		Significant Difference?
Variable	Mean/ Percent	SE	Mean/ Percent	SE		Mean/ Percent	SE	Mean/ Percent	SE	
Percentile (Pretest)	19.32	2.01	42.36	1.59	***	19.48	2.05	19.87	2.52	n.s.
Hispanic	83%	5%	59%	3%	**	85%	5%	80%	6%	n.s.
Other Ethnicity[19]	6%	4%	13%	2%	n.s.	7%	4%	11%	5%	n.s.
English Learner	53%	7%	36%	3%	*	54%	7%	61%	7%	n.s.
Special Ed Student	2%	2%	0%	0%	n.s.	0%	0%	2%	2%	***
Wright Charter[20]	100%	0%	100%	0%	n.s.	100%	0%	100%	0%	n.s.

*$p < .05$. **$p < .01$. ***$p < .001$.

Note: SE = Standard error of mean/percent. n.s. = Nonsignificant.

The sample includes only students who took the STAR360 mathematics assessment (the most common mathematics assessment in the data). Students who took the MI assessment are excluded (N = 40 treatment cases with complete data).

The significant difference between treatment and comparison group for special education student should be interpreted with caution because there was only one student with special education status in the matched sample.

[18] Tests for statistical significance were conducted using regressions that mirror the conclusions of a two-tailed independent samples t-tests for each variable.

[19] White race/ethnicity was not included in the matching since it would be redundant with the Other ethnicity variable, but it served as the reference category.

[20] All students were in the Wright Charter school.

Table 8

Summary of Propensity Score Matching Results for Mathematics Tutoring Program Participants (Spring 2019)

	UNMATCHED Sample					1:1 MATCHED Sample				
	Treatment (N=43)		Comparison (N=252)		Significant Difference? [21]	Treatment (N=43)		Comparison (N=43)		Significant Difference?
Variable	Mean/ Percent	SE	Mean/ Percent	SE		Mean/ Percent	SE	Mean/ Percent	SE	
Percentile (Pretest)	23.17	2.38	52.08	1.74	***	23.17	2.39	23.77	2.81	n.s.
Hispanic	81%	6%	60%	3%	**	81%	6%	84%	6%	n.s.
Other Ethnicity[22]	7%	4%	12%	2%	n.s.	7%	4%	7%	4%	n.s.
English Learner	53%	8%	36%	3%	*	53%	8%	40%	7%	n.s.
Special Ed Student	16%	6%	17%	2%	n.s.	16%	6%	14%	5%	n.s.
Wright Charter[23]	100%	0%	100%	0%	n.s.	100%	0%	100%	0%	n.s.

*$p < .05$. **$p < .01$. ***$p < .001$.

Note: SE = Standard error of mean/percent. n.s. = Nonsignificant.

The sample includes only students who took the STAR360 mathematics assessment (the most common mathematics assessment in the data). Students who took the MI assessment was excluded (N = 57 treatment cases with complete data).

[21] Tests for statistical significance were conducted using regressions that mirror the conclusions of a two-tailed independent samples t-tests for each variable.

[22] White race/ethnicity was not included in the matching since it would be redundant with the Other ethnicity variable, but it served as the reference category.

[23] All students were in the Wright Charter school.

Table 9

Summary of Propensity Score Matching Results for Mathematics Tutoring Program Participants (Fall 2019)

	UNMATCHED Sample					1:1 MATCHED Sample				
	Treatment (N=62)		Comparison (N=606)		Significant Difference? [24]	Treatment (N=62)		Comparison (N=62)		Significant Difference?
Variable	Mean/ Percent	SE	Mean/ Percent	SE		Mean/ Percent	SE	Mean/ Percent	SE	
Percentile (Pretest)	24.69	1.79	47.79	1.10	***	24.69	1.78	25.68	2.08	n.s.
Hispanic	81%	5%	69%	2%	n.s.	81%	5%	87%	4%	n.s.
Other Ethnicity[25]	8%	3%	17%	2%	n.s.	8%	3%	8%	3%	n.s.
English Learner	37%	6%	35%	2%	n.s.	37%	6%	37%	6%	n.s.
Special Ed Student	13%	4%	18%	2%	n.s.	13%	4%	10%	4%	n.s.
RL Stevens[26]	31%	6%	52%	1%	**	31%	4%	34%	4%	n.s.

*$p < .05$. **$p < .01$. ***$p < .001$.

Note: SE = Standard error of mean/percent. n.s. = Nonsignificant.

The sample includes only students who took the STAR360 mathematics assessment (the most common mathematics assessment in the data). Students who took the MI assessment was excluded (there were no treatment cases).

[24] Tests for statistical significance were conducted using regressions that mirror the conclusions of a two-tailed independent samples t-tests for each variable.

[25] White race/ethnicity was not included in the matching since it would be redundant with the Other ethnicity variable, but it served as the reference category.

[26] Wright Charter school variable was not included in the matching since it would be redundant with RL Stevens school variable, but it served as the reference category.

Table 10

Summary of Propensity Score Matching Results for Mentoring Program Participants

	UNMATCHED Sample					1:1 MATCHED Sample				
	Treatment (N=57)		Comparison (N=409)		Significant Difference? [27]	Treatment (N=57)		Comparison (N=57)		Significant Difference?
Variable	Mean/ Percent	SE	Mean/ Percent	SE		Mean/ Percent	SE	Mean/ Percent	SE	
Total SDQ Score (Pretest)	19.98	0.84	20.22	0.27	n.s.	19.98	0.84	19.84	0.69	n.s.
Hispanic	68%	6%	75%	2%	n.s.	68%	6%	60%	7%	n.s.
Other Ethnicity[28]	26%	6%	10%	1%	***	26%	6%	14%	5%	n.s.
Female	30%	6%	53%	2%	**	30%	6%	33%	6%	n.s.
English Learner	33%	6%	30%	2%	n.s.	33%	6%	25%	6%	n.s.
Special Ed Student	35%	6%	18%	2%	**	35%	6%	39%	6%	n.s.
Harvest MS	35%	6%	19%	1%	**	35%	4%	28%	4%	n.s.
Silverado MS[29]	23%	5%	55%	1%	***	23%	4%	21%	4%	n.s.

*$p < .05$. **$p < .01$. ***$p < .001$.

Note: SE = Standard error of mean/percent. n.s. = Nonsignificant.

[27] Tests for statistical significance were conducted using regressions that mirror the conclusions of a two-tailed independent samples t-tests for each variable.

[28] White race/ethnicity was not included in the matching since it would be redundant with the Other ethnicity variable, but it served as the reference category.

[29] Redwood MS school variable was not included in the matching since it would be redundant with the other school variables, but it served as the reference category.

Table 11

Summary of Logistic Regression Results Predicting STAR360 Literacy Performance Level Improvement from Pre-test to Post-test

	Fall 2018			Spring 2019			Fall 2019		
	Matched Students (N = 212)			Matched Students (N = 408)			Matched Students (N = 446)		
Predictor	*B*	*SE B*	e^B	*B*	*SE B*	e^B	*B*	*SE B*	e^B
Any Literacy Tutoring	-0.45	0.41	0.64	-0.27	0.29	0.76	-0.10	0.24	0.91
Grade	-0.66	0.21	0.52**	-0.39	0.11	0.67**	-0.20	0.08	0.82*
Hispanic	-1.53	0.96	0.22	-0.44	0.57	0.64	-0.54	0.55	0.59
Other Ethnicity	-15.01	1.02	<0.001***	1.30	0.74	3.67	-0.37	0.63	0.69
English Language Learner	-0.05	0.67	0.95	0.09	0.42	1.1	-0.67	0.31	0.51*
Special Education Status	-1.69	1.59	0.19	1.04	1.11	2.82	-0.67	0.53	0.51
Bellevue ES	0.17	0.67	1.19	0.18	0.38	1.20	0.03	0.46	1.03
Meadow ES	0.35	0.89	1.42	-0.70	0.60	0.50	0.13	0.38	1.14
RL Stevens	1.46	0.76	4.32	-0.27	0.64	0.76	0.20	0.58	1.22
Taylor Mountain ES	0.13	0.84	1.14	-1.23	0.54	0.29*	0.69	0.37	2.00
Wright Charter	1.46	0.91	4.31	0.62	0.60	1.86	-0.56	0.61	0.57
Intercept	2.06	1.29	7.87	0.33	0.74	1.40	0.96	0.85	2.60
χ^2		3.8			2.97			2.37	
df		11			11			11	
% Improved		19.81%			17.65%			23.77%	

*$p < .05$. **$p < .01$. ***$p < .001$.

Note: e^B = exponentiated *B*. All variables are coded with 0=No and 1=Yes.

White race/ethnicity and Kawana school are reference categories and are not included in the model.

Table 12.
Summary of Linear Regression Results Predicting STAR360 Literacy Percentile at Post-test

	Fall 2018			Spring 2019			Fall 2019		
	Matched Students (N = 212)			Matched Students (N = 408)			Matched Students (N = 446)		
Predictor	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Any Literacy Tutoring	-4.78	2.00	-0.13*	-2.39	1.32	-0.06	-2.76	1.55	-0.07
Percentile (Pre-test)	0.94	0.14	0.52***	0.78	0.05	0.68***	0.80	0.10	0.53***
Grade	-3.03	0.82	-0.26***	-2.19	0.50	-0.18***	-1.33	0.48	-0.11*
Hispanic	-13.40	7.83	-0.24	-1.70	3.92	-0.03	1.04	4.37	0.02
Other Ethnicity	-15.30	7.45	-0.20*	7.91	4.73	0.09	2.65	4.74	0.03
English Language Learner	2.48	3.50	0.05	0.21	2.28	0.00	-6.50	2.41	-0.15*
Special Education Status	-11.80	6.44	-0.12	5.26	5.06	0.04	-2.18	3.24	-0.04
Bellevue ES	-0.87	2.44	-0.02	-0.29	1.58	-0.01	-0.96	2.82	-0.02
Meadow ES	1.42	4.39	0.02	-3.78	2.39	-0.06	2.23	1.81	0.05
RL Stevens	4.52	3.64	0.08	-1.59	2.73	-0.02	4.09	3.79	0.07
Taylor Mountain ES	-0.12	2.93	0.00	-5.58	1.72	-0.12**	8.97	2.39	0.19***
Wright Charter	4.93	4.40	0.09	0.97	3.35	0.01	-1.40	2.82	-0.03
Intercept	32.86	8.26	0.00	17.44	4.02	0.00***	19.49	6.54	0.00**
R^2		0.42			0.58			0.39	
F		11.91			37.88			14.05	

*$p < .05$. **$p < .01$. ***$p < .001$.
Note: White race/ethnicity and Kawana school are reference categories and are not included in the model.

Table 13.

Summary of Logistic Regression Results Predicting STAR360 Math Performance Level Improvement from Pre-test to Post-test

	Fall 2018			Spring 2019			Fall 2019		
	Matched Students (N = 92)			Matched Students (N = 86)			Matched Students (N = 124)		
Predictor	*B*	*SE B*	e^B	*B*	*SE B*	e^B	*B*	*SE B*	e^B
Any Math Tutoring	-0.92	0.50	0.40	-0.81	0.58	0.44	0.01	0.40	1.01
Grade	-0.29	0.15	0.75	0.06	0.16	1.07	-0.39	0.13	0.68**
Hispanic	0.26	0.87	1.30	0.89	1.18	2.44	-0.34	0.82	0.71
Other Ethnicity	-0.11	1.18	0.90	0.51	1.56	1.67	-1.32	1.12	0.27
English Language Learner	-0.73	0.61	0.48	-0.10	0.68	0.90	-0.64	0.44	0.53
Intercept	1.33	1.03	3.78	-2.09	1.56	0.12	2.14	1.04	8.53*
χ^2		1.90			0.71			2.91	
df		5			5			5	
% Improved		31.52%			20.93%			41.94	

*$p < .05$. **$p < .01$. ***$p < .001$.

Note: e^B = exponentiated *B*. All variables are coded with 0=No and 1=Yes.

White race/ethnicity and Kawana school are reference categories and are not included in the model.

Table 14

Summary of Linear Regression Results Predicting STAR360 Math Percentile at Post-test

	Fall 2018			Spring 2019			Fall 2019		
	Matched Students (N = 92)			Matched Students (N = 86)			Matched Students (N = 124)		
Predictor	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Any Math Tutoring	-12.64	3.90	-0.27**	-5.21	3.49	-0.12	0.65	3.25	0.02
Percentile (Pre-test)	0.78	0.14	0.52***	0.80	0.10	0.65***	0.74	0.10	0.54***
Grade	-2.83	1.30	-0.21*	-1.25	0.92	-0.11	-3.86	0.96	-0.33***
Hispanic	8.47	6.41	0.14	2.58	6.67	0.05	0.42	6.49	0.01
Other Ethnicity	5.07	9.66	0.06	-2.51	7.17	-0.03	-8.99	7.41	-0.12
English Language Learner	-1.82	5.01	-0.04	-3.70	3.89	-0.09	-4.22	3.67	-0.10
Intercept	28.80	9.24	0.00**	15.47	9.35	0.00	40.29	7.91	0.00***
R^2		0.38			0.49			0.34	
F		9.84			13.72			13.86	

*$p < .05$. **$p < .01$. ***$p < .001$.

Note: White race/ethnicity and Kawana school are reference categories and are not included in the model.

Table 15
Summary of Linear Regression Results Predicting SDQ Total Score at Post-test

	Matched Students (N = 114)		
Predictor	*B*	*SE B*	*β*
Any Mentoring	0.19	0.91	0.02
Total Score (Pre-test)	0.63	0.08	0.61***
Grade	0.10	0.65	0.01
Hispanic	-1.00	1.41	-0.08
Other Ethnicity	-2.11	1.80	-0.14
Female	0.28	1.01	0.02
English Language Learner	-1.64	1.04	-0.12
Special Education Status	0.25	1.15	0.02
Harvest MS	-2.03	1.27	-0.16
Silverado MS	-0.63	1.08	-0.04
Intercept	8.49	4.64	0.00
R^2		0.45	
F		14.09	

*$p < .05$. **$p < .01$. ***$p < .001$.
Note: White race/ethnicity and Redwood school are reference categories and are not included in the model.