Changing the Odds:
Student Achievement after Introduction of a Middle School Math Intervention

By Julian R. Betts, Andrew C. Zau, Karen Volz Bachofer and Dina Polichar
San Diego Education Research Alliance at UCSD,
Department of Economics,
UC San Diego
Sandera.ucsd.edu

This Draft: June 29, 2021
First Draft: 9/29/2020

Acknowledgements: The work described in this paper has been supported by grant R305H150028 from the U.S. Department of Education under its Continuous Improvement grant competition. All statements and opinions expressed in this paper are the sole responsibility of the authors. We are grateful to our many research partners on the Changing the Odds project including Ronald Rode, Amanda Datnow, Kimberly Samaniego, Ann Trescott, and Hayley Weddle, Leah Baylon and to our partners who implemented key parts of the reform, Kira Rua and Andrea Barraugh, for helping to make this project possible.

Abstract

The paper evaluates math performance at four high-need middle schools during a four-year intervention, which was designed to help math teachers diagnose students' areas of need and to design lesson plans responsive to those needs. Before the intervention began the evaluation team pre-selected four middle schools as comparison schools by matching based on demographics and achievement. A difference-in-difference analysis finds a significant increase of about 0.11 standard deviation in test scores per year for students in the program schools. Supplementary event study and synthetic control analyses to detect year-by-year effects lack precision but are suggestive of increasing gains over time.

Introduction

Performance of secondary school students in math courses is one of the best predictors of college outcomes and early career wages in the United States (Rose and Betts, 2004). At the same time, the U.S. ranks poorly on international math tests. For instance, the 2015 Program for International Student Achievement (PISA), a test of 15-year old students conducted by the Organization for Economic Cooperation and Development (OECD), ranked the U.S. 40th among participating education systems (using the mean score), which was significantly below the OECD average. The U.S., with a mean score of 470, ranked far below the average score in the top-ranked country, Singapore, at 564 (OECD, 2016, Table I.5.1). Similarly, the 2015 Trends in International Mathematics and Science Study (TIMSS) illustrate that U.S. grade 8 math scores were significantly lower than scores for nine other countries (Provasnik et al., 2016, Figure 1b.).

National Assessment of Education Progress (NAEP) data consistently show that U.S. students struggle with math and that their performance worsens in middle school, with proficiency rates dropping sharply between grades 4 and 8. (See e.g. NCES, 2014 a,b,c.) Figure 1 shows that this drop-off in proficiency applies to data for the nation, California and the California district hosting the program which this paper discusses, San Diego Unified School District.

This study provides a quantitative evaluation of a continuous improvement project, Changing the Odds, expressly designed to improve math achievement in high-need middle schools. The project was funded as a Continuous Improvement grant by the U.S. Department of Education. The key elements of the program, which lasted four years, included diagnostic mathematics assessments to identify topics that students had yet to master, peer coaching of math teachers to help them design lessons that address these needs, and support for the Professional Learning Communities (PLCs) within each school.

The intervention itself was co-designed by policymakers and subject matter specialists in the host district, San Diego Unified School District, and a university team at the University of California San Diego with expertise in diverse areas, including data driven decision making, diagnostic testing, and evaluation of educational reforms.

Several bodies of literature underlie the program. First, the literature on data driven decision making (DDDM) studies how teachers examine data about student performance in order to inform instructional changes that lead to improved learning (Datnow & Park, 2014). Although studies have shown that appropriate use of student data can improve math outcomes (Feldman & Tung, 2001; Light et al., 2005), diagnostic information on student strengths and weaknesses by itself does not guarantee improved outcomes. Teachers need to know how to interpret the data, and how to help their students improve.

The second body of literature motivating the intervention examines how teachers work together in Professional Learning Communities (PLCs) to improve their teaching (Honig &

Venkateswaran, 2012; Marsh, 2012; Means, Padilla, & Gallagher, 2010). PLCs offer an opportunity for many types of collaboration, including jointly examining student work and test results to make plans to support student learning. (See e.g. Blanc et al., 2010; Cosner, 2011; Datnow & Park, 2019; White & Anderson, 2011.) Again, as with DDDM, the existence of a PLC at a school does not by itself ensure improved teaching. For example, Daly (2012) points out that PLCs consisting of teachers who have less expertise can lower teaching effectiveness due to misinterpretation of data by the group, and promulgation of ineffective teaching practices.

The third body of literature which informed the Changing the Odds program is the work on Continuous Improvement. Continuous Improvement (CI) research is based on the idea of quite frequent cycles of "Plan-Do-Study-Act" in which an initial plan is developed and implemented (Plan-Do) to be quickly followed by data gathering and analysis, reflection, and fine-tuning of the original plan (Study-Act) (Deming, 1986). Under this approach, teachers can use annual measures of performance to change their instruction. In addition a potentially more valuable approach is for teachers to use mini-cycles in which they design unit plans and the lesson plans within each unit, in light of what they have learned about their students' recent progress. Regular formative assessments, including exit slips to gauge student understanding at the end of a class, play a key role in helping teachers to study and then act by revising their teaching.

The teacher-centric intervention evaluated in this paper encouraged middle school math teachers to use data to inform instruction. A full-time resource teacher worked with the math teachers to build lessons that responded to data on students' needs. The work occurred in PLC meetings, in regular within-grade across-school unit planning days, through co-teaching by math teachers and the peer coach, and through an annual summer professional development event.

This paper examines the math achievement of students at the schools participating in this program over the course of four years. Difference-in-difference models show a statistically significant positive effect on math achievement in the Changing the Odds (CTO) schools. Supplementary analyses suggest that the reform had an increasing impact over time, which meshes with the increasing frequency of interaction between the peer coach and the schools' math teachers over time.

The next section describes the intervention in more detail, and also lays out the state testing programs that existed around the time of the intervention, and implications for design of the evaluation.

Background on the Intervention

The CTO program was implemented in four high-need middle schools in the San Diego Unified School District (SDUSD), between the 2015-16 and 2018-19 school years.

The approach has the following components:

- DDDM, with regular provision to teachers of well-organized data provided by the university team and SDUSD, to pinpoint areas in which students struggle in math. These include diagnostic assessments at the beginning of each semester and year-end diagnostic and summative assessments. Teachers were also encouraged (and trained) to use formative assessment data they gathered in their own classrooms.
- PLCs (grade or department level team meetings) in which math teachers at a school worked together to use assessment data to monitor students' progress, to improve the quality of math instruction, and to plan collaboratively. The resource teacher focused on innovative approaches to inspiring students to participate and to learn, along the lines suggested by authors such as Boaler (2015). All schools had PLCs in place when the project began, and the project aimed to build upon this pre-existing structure.
- Professional development support in the form of:

1) a grant-funded project resource teacher who worked with math teachers in the four middle schools;
2) curricular materials and professional development;
3) a professional development consultant who was hired in year two to support the work of the resource teacher and to co-plan activities for teachers;
4) opportunities for cross-school collaboration for math teachers at the four schools to engage in unit planning; and
5) a two- to three-day Summer Math Summit each year for teachers and principals from the four schools, along with district administrators, the CTO resource teacher and the research team, to share lessons learned, to analyze end-of-year assessment data, to analyze student work for understandings and sources of misunderstanding (Ashlock, 2010), and to refine plans for the upcoming year;
6) assessment support from Mathematics Diagnostic Testing Project (MDTP) staff in how to make the best use of MDTP diagnostic tests and instructional support materials. (The tests are described more fully below.)
7) ongoing feedback from the UCSD research team to inform the process of continuous improvement.

These intersecting project networks allowed the team to engage in improvement research (Bryk et al., 2011).

Several of the district and university partners, as former teachers, had observed that resource teachers were useful supporters of classroom teachers in the host district, but that spreading a resource teacher's time and talents across too many schools was unlikely to lead to large gains in student achievement. This led to the decision, while in the planning stage, that the resource teacher to be hired as part of the project should devote her time to four schools.

Once the intervention was underway, near the start of each school year students in the four schools were given diagnostic tests designed by the Mathematics Diagnostic Testing Project (MDTP). The MDTP is a joint program of the California State University and the University of California. It offers free diagnostic testing to math teachers throughout California. The MDTP is designed to allow teachers to obtain diagnostic testing of their students, with detailed feedback. In a series of "readiness" tests, which test the skills prerequisite for good performance in an upcoming course, the MDTP uses a multiple

choice test with carefully designed distractor answers. Teachers receive, almost instantly, feedback on average student performance in each of several specific areas, and information on what the most common wrong answers were, and what student misconception probably led to the student choosing each wrong answer. Students and their families receive feedback on areas in which students have adequate understanding and areas in which they need to improve. Teachers can use the online response data to group students by their answers to specific questions, and to develop plans to address common misperceptions.

Roughly a decade before the current study, for several years the host district made MDTP testing mandatory in certain grades. Betts, Hahn and Zau (2017) find that this mandated use of MDTP led to improvements in subsequent math achievement for the affected grades.

Findings from the MDTP tests were used to help the resource teacher and MDTP staff plan activities for the annual Summer Math Summit. For instance, an early finding from the MDTP test results was that many students had memorized rules about operations on fractions, but lacked deep conceptual understanding. This often led to confusion about when to apply which rules. Therefore, one Summer Math Summit devoted time to planning on instructional activities that could both correct and deepen students' understanding of operations on fractions.

Continuous improvement applied to the mini-cycles in which teachers would plan and then implement lessons, use formative assessments to assess how well students had mastered the material, and finally, adjust subsequent lessons. These continuous improvement practices clearly led to changes over time in how the teachers worked together and with the resource teacher. The two most important adjustments came after the first year. At the first Summer Math Summit in August 2015, teachers expressed enthusiasm about working together across schools in within-grade groups, and asked for more of these events. The UCSD team learned after several failed attempts in year 1 that bringing teachers together across schools for unit planning after school was not practical, due both to variations in the schools' bell times but also the parenting responsibilities that many of the teachers had after school. No events of this type occurred in year 1. In years 2, 3 and 4, a series of all-day unit planning events were conducted, within grade and across the four schools. A second reform that began in year 2 (2016-17) involved a change in how the resource teacher was hosted in the district. In year 1, like other district resource teachers, the project resource teacher was provided office space in the district headquarters. In later years, the resource teacher was instead hosted in the four CTO schools on a rotating basis, leading to far more interactions with the math teachers. A third example of changes over time is that in year 3 and especially year 4, on several occasions teachers in the current host school worked with the resource teacher to design a lesson plan, then took turns co-teaching the material in each other's classes, and refined the lesson plan at each iteration.

Empirical Approach

The empirical approach to studying whether student math achievement changed after the introduction of the program takes into account the timing of both the CTO program and the transition in California from one student testing system to another. The first column of Table 1 shows the four years during which the CTO program was in effect. The next two columns show the two standardized tests used in California in spring 2009 and later years. We start in 2008-2009 because one of the four CTO schools was reorganized in that year, meaning that comparing trends across schools in years before that would not be appropriate. (To save space, henceforth we refer to school years by the year in which spring semester took place.)

The second and third columns show the years in which the California Standards Test (CST) and the Smarter Balanced test were used in California. As the table shows, the CST test was used up to 2013. The state mandated no test in 2014 and then in 2015 introduced its version of the Smarter Balanced test. A second complication is that the CST math test was grade-specific up to grade 7, but in grade 8 and higher, when students took different math courses, students also took different CST tests focused on the course taken. This makes it very difficult to compare the achievement level of students in grade 8 who took different CST tests.

One quasi-experimental approach to evaluating the impact of the CTO program is to use a difference-in-difference approach, comparing achievement before and after the CTO program began in the four CTO schools and comparison schools. We pre-committed when designing the intervention to compare to four other middle schools that we matched in 2014 based on 2013 test scores and demographics. Appendix Table 1 shows the variables upon which we matched. (The reported differences are exact, but we round the means so as not to reveal the identity of the eight schools inadvertently.)

Several issues must be addressed when using CST math scores in the earlier years and Smarter Balanced test scores in later years. The first issue is that we must restrict the analysis to math performance in grades 6 and 7 because there is no consistent way to compare CST scores among grade 8 students.

Second, we must put the two tests on a common metric, so we converted all test scores to Z-scores, using the district average and standard deviation in each year and grade. Thus a math score of -0.2 means that the student scored 0.2 of a standard deviation below the district mean for the given grade and year. This standardization, however, does not guarantee that the two tests are measuring math achievement in a similar way. Given that no test was given in 2014, we assessed the relationship between the two tests as follows. First, we correlated the Smarter Balanced test scores of grade 6 and 7 students in the district in 2015 with the same students' scores two years earlier on the math CST. The correlation was quite high, at 0.76. We repeated this calculation using CST scores in the same grades in 2013 relative to the students' 2011 CST scores, and the correlation was slightly lower, at 0.69. The implication is that comparing Z-scores from the Smarter

Balanced and CST math tests in grades 6 and 7 makes sense in that students' rankings two years apart are quite stable. This is true regardless of whether we use only CST data or whether we compare the first year of the Smarter Balanced tests to the same students' scores on the CST two years earlier.

Figures 2 and 3 show the grade 6 and 7 Smarter Balanced Z-scores in 2015 plotted against the same students' 2013 CST scores, and the 2013 CST grade 6 and 7 scores plotted against the same students' 2011 CST scores. Figure 2 suggests a clear, almost linear, relationship between the new Smarter Balanced test scores and the CST scores from two years earlier. One distinction is that the CST scores become somewhat granular at the upper end, while the Smarter Balanced scores show a more continuous distribution across the range. This likely reflects the use of computer-adaptive testing in Smarter Balanced, which can feed a more advanced set of questions to students who perform well on earlier questions on the test, reducing ceiling effects. Apart from this, the strong relation between the two tests is quite apparent. Figure 3, which plots CST scores against twice lagged CST scores, is very similar. This suggests that comparing Z-scores from the two tests is meaningful.

Figure 2 does show a slight curvature in the relation between Smarter Balanced math scores and the twice lagged math CST scores, for those with CST scores roughly two standard deviations above the district mean. But because the CTO schools are low-performing schools, as are the four comparison schools, very few students in those schools had test scores in that range. Figure 4 illustrates this, using grade 6 and 7 math CST scores in 2013 for these eight schools. The 95th percentile math Z-score for this sample was only 1.32, and the mean was -0.32.

To assess the influence of the CTO program on student outcomes, we employ a difference-in-difference approach. These models combine CST and Smarter Balanced test scores. The main specification is as follows. Let S_{igst} denote the math Z-score of student i in grade g and school s at time t. To take into account past learning, we take a value-added approach, modeling the student's annual change in math achievement ΔS_{igst}. We model this change as a function of year and grade dummy variables, a vector of student characteristics X_{it}, and a dummy variable to capture any differences in average learning over all the years observed between CTO and comparison schools. The most important regressor, CTO*$TREAT_{ist}$, captures the difference-in-difference, that is, the difference in test score changes at CTO schools after the treatment began, relative to the same before-after difference observed in comparison schools:

$$\text{(1a)} \quad \Delta S_{igst} = \alpha_t + \beta_g + \sum_{j=1}^{J} X_{it,j}\,\varphi_j \quad + \theta CTO_{ist} + \pi CTO_{ist} * TREAT_{ist} + \varepsilon_{igst}$$

In this equation, the Greek letters indicate coefficients to be estimated, except for the final term, ε_{igst}, which is an error term. This model allows for a fixed difference in the outcome between the CTO schools and the comparison schools, with θ measuring this difference. This is a useful measure because it indicates whether the treatment and comparison schools had similar changes in test scores in the pre-treatment years. We also

estimate the more traditional diff-in-diff model with fixed effects for each school, indicated by the λ terms below:

$$(1b) \quad \Delta S_{igst} = \alpha_t + \beta_g + \lambda_s + \sum_{j=1}^{J} X_{it,j}\,\varphi_j \quad + \pi CTO_{ist} * TREAT_{ist} + \varepsilon_{igst}$$

We estimate these models while clustering the error terms to allow for arbitrary correlations between all of the observations for a given school s. Because we have only four treatment and four comparison schools, and clustered standard errors can become unreliable with a small number of clusters, we use the wild bootstrap approach of Cameron, Gelbach and Miller (2008). In this approach we choose the wild weight using the Webb (2014) six-point distribution, which is preferable when the number of clusters is less than 10 or 12 (see e.g. Roodman, MacKinnon, Nielsen and Webb, 2019).

Because it is also worthwhile to see how CTO and comparison schools fared in each year of the CTO program, 2016-2019, we also estimate event study models. These are variants of (1a) and (1b) that replace the CTO_{ist}*$TREAT_{ist}$ indicator with terms that interact CTO_{ist} with dummy variables for each year except 2013, the last year of the CST test and the last year in which we can observe gains in achievement prior to the start of the intervention, so that 2013 serves as the base year against which we compare the CTO-comparison difference in other years.

As mentioned previously, the researcher-practitioner team designed the intervention to focus its resources on four schools. We reiterate this point here because the relatively small sample of schools means that only moderate or large effects were likely to be detectable. This is especially true for the event study which estimates the difference in outcomes between the CTO and comparison schools separately for each school year.

In a robustness check that restricts the test used to Smarter Balanced only, while still allowing a difference-in-difference comparison, we also model the levels of test scores, using the Smarter Balanced test scores available from 2015 through 2019. (By modeling achievement level, rather than changes, we obtain one pre-treatment year.) For this approach we use the synthetic control approach (Abadie and Gardeazabal, 2003, Abadie Diamond and Hainmueller, 2010). These authors' method uses pre-treatment years to create a synthetic control, that is, a weighted average of non-treated units, by matching based on one or more variables that consist of a mix of linear combinations of the outcome in the pre-treatment periods and, optionally, predictors of those pretreatment outcomes. In this method if the unit (school) we label as 1 is treated and there are J potential units that could serve as matches, the algorithm produces a JX1 vector of non-negative weights W that sum to 1. The weights are chosen to minimize

$$\sqrt{(X_1 - X_o W)' V (X_1 - X_o W)}$$

where X_0 is a matrix of the variables upon which matching isto be performed, for each of J potential comparison schools, and X_1 is a vector of the same variables for the treated unit. V is a symmetric and positive semidefinite matrix, which is chosen so that the weights chosen, W, lead to the smallest possible Root Mean Squared Prediction Error (RMSPE) for the outcome variable in the pre-treatment period. In the second step, the entire path of

outcomes is calculated for the synthetic control unit, which is calculated as the weighted average of the potential comparison units, and the effect of the intervention is estimated as the difference between the actual outcomes in each treatment year for the treated unit and the synthetic control.

Inference is based on a permutation technique, where the non-treated units are one at a time assumed to be treated, with the remaining comparison units being used to create a synthetic control. Suppose that the difference between the outcome for the actually treated unit 1 and the weighted outcome for its synthetic control in year b of treatment is $\hat{a}_{1,b}$. One calculates the two-sided p-value for this estimated effect as the proportion of the times that this estimated effect, in absolute value, is less than or equal to the absolute value of the corresponding estimate for the J placebo estimates. Below, the superscript PL indicates an estimate for one of the J placebos:

(2) $$p = \Pr\left(\left|\hat{a}_{1,b}^{PL}\right| \geq \left|\hat{a}_{1,b}\right|\right) = \frac{\sum_{j \neq 1} 1\left(\left|\hat{a}_{j,b}\right| \geq \left|\hat{a}_{1,b}\right|\right)}{J}$$

We use the adaptation of this method introduced by Cavallo, Galiani, Noy and Pantano (2013) for cases in which there is more than one unit treated, as in our case, where there are four schools. Their approach is to calculate all possible placebo average effects by picking a single placebo estimate for each treated unit and then averaging across the four treated units. This is repeated for all possible combinations. For example, if there were two treated units with one having 20 possible comparison units and the second having 30 possible comparison units, the number of combinations of averages of placebo effects would be 20 X 30 = 600. The p-value for the average treatment effect is then the proportion of cases in which the actual average treatment effect is less than or equal to the averaged placebo effect, both calculated as absolute values.

In our case we choose a single predictor, which is the 2015 mean math Z-score, which produces a very good pre-treatment fit. This simple approach of matching based on the key outcome is an approach that dispels concerns that one can manipulate synthetic control findings by choosing repeatedly from a very wide set of covariates.

One practical issue with inference for the synthetic control method is that if some of the placebo units have very poorly fitted synthetic controls in the pre-treatment period, the estimated impacts for these placebos could be very large. This would make our inference too conservative. Following , Abadie Diamond and Hainmueller (2010) and Cavallo, Galiani, Noy and Pantano (2013), we restrict placebos to those with RMSPE less than a certain multiple of the RMSPE for the CTO schools. We report results for the multiple set to 2, but we report how sensitive inferences change with different multiples.

Results

Table 2 shows the difference-in-difference estimates for three variations of (1a) and (1b). The first three columns show results for model (1a) with no student demographic variables added (column 1), with indicators for female students and for four of five racial/ethnic categories (White, Black, Hispanic, Asian and Other) (column 2), and with these measures plus indicators for whether the student's parents had a high school diploma or less, and for whether the student was ever an English Learner (EL) (column 3). We use the "ever" EL indicator rather than current EL status because the latter could be an endogenous regressor. Columns 4 to 6 repeat these specifications but replace the CTO indicator with dummy variables for each school, as in model (1b).

The results are similar across specifications. The CTO schools break significantly from the comparison school trend in the treatment years. The models indicate annual changes in test scores in the CTO schools that are about 0.13 and 0.11 standard deviations above the changes observed in the comparison schools in the estimate for models (1a) and (1b) respectively. (These differences are measured as the proportion of a standard deviation of the levels of test scores observed district-wide in the given school and year.) These are meaningful differences.

The event study models extend (1a) and (1b) by estimating differences in the dependent variable, math score changes, between CTO and comparison schools for each year. We estimated the same six models shown in Table 2 for the diff-in-diff models. Typically, when doing an event study one sets the comparison year to the year before the intervention starts. However, because the CST ended in 2013 and the Smarter Balanced test did not start until 2015, we have test score changes for 2009 through 2013 and 2016 through 2019. We make 2013 the comparison year. Table 3 shows results. The table shows coefficients by year for the event-study versions of (1a) and (1b) that include controls for gender and race/ethnicity. The other four models produced results that were very similar. Because wild-bootstrap confidence intervals need not be symmetric, we show the 95% confidence intervals to the right of the corresponding coefficients. Figure 5 graphs the results along with the confidence intervals.

When we disaggregate the contrast between CTO-comparison school into year-by-year calculations in this way, we lose precision. None of the CTO-year coefficients is significant. However, the coefficients in all four of the treatment years are positive. Apart from a slight drop in year 3 (2018), the estimated treatment effects are increasing during each of the four years of the CTO program. Although the results are suggestive of a growing positive impact in the CTO schools, the lack of precision leads to a conclusion of no significant effects. For example, in evaluating the wild-bootstrap p-values for the coefficients on the 2016 through 2019 CTO coefficients, across 6 specifications (three sets of demographic regressors for event study versions of (1a) and (1b)), the smallest p-values are for the CTO effect in 2019; these p-values varied between 0.123 and 0.138.

We now turn to results on various robustness checks.

First, the choice of comparison schools was made in 2014 based on data available at that time. It could be that better matches could be found, especially given that data on schools in 2015, the year before the intervention began, were not available at the time of the match. Second, instead of choosing four comparison schools, as we had done in 2014 before the project started, one could use a more sophisticated approach such as synthetic control to derive a weighted average of non-CTO schools for the comparison group. Third, in spite of evidence above that the CST and Smarter Balanced tests are measuring math achievement in similar ways, it would be useful to see if similar results emerge without combining the two tests across years. Fourth, on a closely related issue, the value-added models described above examine changes in students' test scores for grades 6 and 7 only, even though all the middle schools enrolled students in grades 6 through 8. The lack of comparable CST scores in grade 8 necessitated this restriction.

A robustness check that would respond to these all four of these concerns is to estimate a synthetic control model for 2015 through 2019, examining levels of Smarter Balanced scores. We use test score levels, rather than changes, for these models because synthetic control models require at least on pre-treatment year, and Smarter Balanced testing debuted in 2015. But this approach maintains the logic of the difference-in-difference approach by comparing CTO and comparison schools before and after the implementation of the CTO program in 2016. One advantage of this approach is that it does not require us to combine Z-scores from different tests. A second advantage is that we can match schools based on the 2015 Smarter Balanced data, from the year before the CTO program started. A third advantage is that by using only Smarter Balanced testing years, we can now include grade 8 students with the grade 6 and 7 students used earlier. Indeed, we can estimate the synthetic control model twice, once with grade 6 through 8 records, and once with grade 6 and 7 records. Comparing these results can provide insights into whether our main analyses described above are likely to have differed substantially if we had been able to include grade 8 records in the years when the CST test was used. Against these two major advantages, the obvious disadvantage is that we cannot estimate value-added models because there is no pre-treatment change in test scores available that uses only Smarter Balanced scores.

Figures 6 and 7 show the results of the synthetic control models for the average Z-score levels in the four CTO schools relative to a synthetic control group. Figure 6 uses the average of grade 6 and 7 Z-scores to conform with the preceding diff-in-diff and event study models. Figure 7 instead averages over all students, that is, all in grades 6 through 8. Table 4 shows the estimated effects by year and the p-values based on the wild bootstrap.

To obtain a synthetic control group that matched the CTO schools well in the single pre-treatment period, shown in the figures as period 0, we used one predictor, which was the actual mean Z-scores in that period. We restricted potential comparison schools to those whose pre-treatment RMSPE was no more than twice that for the actual CTO schools. When we increased these limits to 5 or 10, p-values changed by small amounts. No

estimates crossed the 1% or 5% significance level thresholds relative to those shown in Table 4.

The results are roughly similar between the samples with and without the grade 8 students, but larger in one sense when the grade 8 students are included. Both models suggest no significant differences in year 1 (2016) or year 3, but a highly statistically significant difference by 2019 of 0.141 standard deviations in the grade 6 and 7 model and 0.145 standard deviations in the grade 6 through 8 model. The major difference is year 2, where the effect is twice as high when grade 8 students are included. Moreover, the year 2 effect is highly significant in the grade 6 to 8 model but not significant in the grade 6 to 7 model. The figures indicate that the difference in the size of the year 2 effect is due both to a slightly increased outcome for the CTO schools, and a much larger drop in achievement in the synthetic comparison school in the grade 6 to 8 model.

The value-added event study effects by year of the CTO program can be compared to the synthetic control estimates that model levels of achievement. Figure 8 re-plots the synthetic control results in Figure 6 by showing the estimated effect size, rather than the outcomes for the treatment and control group. This facilitates comparison with the event study in Figure 5. Both show similar patterns, with an effect near zero initially but with the exception of year 3, effects that become more positive over time. The corresponding effect sizes appear in Tables 3 and 4. While the two figures are not directly comparable, because the event study results in Figure 5 model changes in test scores, and the synthetic control results in Figure 8 model levels of test scores, the patterns showing growth both indicate that the CTO program had little effect in year 1 but grew in impact over time.

We draw two principal conclusions from this analysis. First, the pattern of estimated effects on test-score levels is actually quite close to the event study in which we modeled changes in test scores over many years.

Second, the CTO effects emerged sooner when we include grade 8 students, but by year 4 the effects are similar in magnitude and significance between the grade 6 to 8 and grade 6 to 7 samples. This provides some assurance that the diff-in-diff and event study models earlier in this section, which had to focus on grades 6 and 7 students to allow comparability across CST tests, are probably not influencing the effects to a large degree by dropping grade 8 students.

Estimating Costs of the Program

Although the project did not include a formal cost-benefit analysis, we can estimate costs quite closely by considering the additional labor devoted to the four project schools, plus some expenditures on other elements such as the contract with the consultant hired to help the resource teacher, and spending on items including supplies. The resource teacher worked with roughly 40 math teachers at the four schools, and had a level of experience and a pay grade slightly above the middle of the pay range for teachers with average experience in the district. The researcher-practitioner team devoted considerable

resources to facilitating and observing the reform, but many of these were one-time development costs and costs that arose from the research component of the work. Additional annual costs included payments to teachers for the annual two- to three-day Summer Math Summits, and for substitute teachers when teachers in each grade participated with the resource teacher in a unit planning day. Another cost arose on days when the resource teacher and the regular math teachers would co-plan and co-teach lessons, and then revise the lesson before teaching it to the next class. On these days, the project hired a substitute teacher at the school hosting the resource teacher.

The top panel in Table 5 shows the costs in terms of the number of average teacher days devoted to each of these resources. We use actual teacher days for participating teachers but inflate the days for the full-time resource teacher by 20% to account for this teacher's higher than average qualifications. The first row shows the estimated cost for the resource teacher, in teacher days, while the next four rows show costs in terms of teacher time paid for related to the annual Summer Math Summit, unit planning days, co-teaching, and for the time of teachers participating in professional development related to using the Desmos online learning platform. On top of this, annual expenditures to hire an expert in professional development who coached and supported the resource teacher translates into 19.4 teacher-days a year. The bottom row in the top panel also lists the costs of hiring a consultant to provide the aforementioned professional development one day in year 4 on how to use the Desmos online platform, and for various books and other supplies that the project provided to teachers.

The second panel in Table 5 shows the total cost in various ways. Combining all four CTO schools, total annual average expenditures totaled just over 329 Teacher Days per year, which equals 1.8 teachers on a full time basis. Because four schools participated in the program, the cost per school per year was about 0.45 teacher Full Time Equivalents per year.

In dollar terms the annual cost was just over $200,000 per year or just over $50,000 per school per year. To provide a sense of how much this expenditure was in relative terms, the bottom row of the table expresses this annual expenditure per school as the percentage of the estimated wage and benefit bill for math teachers at the given school. In a typical year, schools employed ten math teachers on average. The CTO expenditures per school represent a 4.5% increase in expenditures beyond our estimate of the average school spending *on math teachers*. Because math teachers are a subset of all the teachers teaching in middle schools, this expenditure is likely to be far below 1% of the total personnel costs for the CTO schools across all subject areas.

Due to housing segregation, high-need students tend to be concentrated in certain schools. Thus, a district that aimed to improve math achievement in a subset of its highest need schools could implement a program similar to CTO without needing to increase personnel cost for math teaching by 4.5% districtwide. For example, if a district chose the schools comprising the third of schools with the greatest student needs for such a program, districtwide the costs would amount to an overall districtwide increase in personnel costs related to math teaching of about 4.5%/3 or 1.5%. This is a modest

amount, and compensatory funds that many states provide through their education finance systems could in theory be tapped to fund similar programs.

This additional cost is modest but not inconsequential. Another way to see this cost is in the light of compensatory education finance reforms. California, like many other states, provides additional funding to districts in proportion to the share of students in the districts who face additional educational challenges, for instance due to students living in poverty, English Learners, students with special education needs, and foster children.

In California the Local Control Funding Formula (LCFF) allots additional funds for districts to spend on these students. [1] The LCFF consists of three components, a base grant based on Average Daily Attendance (ADA), a supplemental grant of about 20% over the base grant which is limited to the proportion of ADA that consists of students who are English learners (EL), students meeting income requirements to receive a free or reduced-price meal (FRPM), foster youth, or any combination of these factors, with an unduplicated count for students meeting more than one criterion. The third tier of funding, known as Concentration Grants, is limited to those districts where the targeted students represent more than 55% of enrollment. The grant is 50 percent of the adjusted base grant per pupil multiplied by ADA and the percentage of targeted pupils exceeding 55 percent of a school district's or charter school's enrollment, divided by 100. In SDUSD, in 2018-19 the Supplemental and Concentration grants divided by total district enrollment equaled $1,023 per pupil and $271 per pupil respectively. The CTO schools had a higher percentage of target students than the district as a whole, but even using the conservative assumption that these funds would be distributed on a per student basis districtwide, without regard to which schools had more higher-need students, the Supplemental and Concentration Grants amount to increases in funding beyond the Base funding of roughly $1 million per CTO school. Seen in this light, the estimated CTO program cost, of $51,974 per participating school per year, seems like it would be quite sustainable in urban districts in California similar to SDUSD, as these are only about 5% of the Supplemental and Concentration grant values. [2]

[1] For more information on the LCFF see information provided by the California Department of Education at https://www.cde.ca.gov/fg/aa/lc/.

[2] In SDUSD, as in other California districts, some of the Supplemental and Concentration funding goes towards central administration needs. But countering that, SDUSD recognizes that students who meet multiple criteria, such as English Learners who are eligible for meal assistance, have greater educational needs than those who are English Learners who are not eligible for meal assistance.. Therefore the district uses duplicated counts of student needs in allocating funds to schools, which effectively increases funding to schools such as the CTO schools where many students meet two or more of the high-need criteria listed earlier.

Conclusion

The Changing the Odds project was designed to help math teachers at high-need middle schools both diagnose students' learning needs and design lesson plans aimed at meeting those needs. This paper examines trends in math learning at these schools compared to a demographically similar set of four comparison schools pre-selected a year before the project activities began.

The paper primarily employs a value-added approach, in which individual students' changes in Z-scores were modelled as a function of student demographics and an indicator for whether the student was attending a CTO school during one of the program years. The difference-in-difference model suggests a positive and significant effect. A version of this model that estimated separate effects by the year of the program was not sufficiently precise to reveal significant differences by the year of the program, but with the exception of a dip in year 3, the remaining years of the intervention showed effects that suggest an increasing effectiveness of Changing the Odds across the four years of the program.

An issue with the above analyses is that they exclude grade 8 test scores due to the lack of comparability of grade 8 test scores between students taking different math courses in most of the pre-treatment years. We supplemented these analyses with a synthetic control model of the level of test scores. This approach is less convincing because it is not a value-added approach that takes into account students' past achievement. But it does bring advantages, including a much broader search for comparison schools, and a method to look at achievement in grades 6 through 8 rather than just grades 6 and 7. These analyses found similar results to the value-added approach, with rising estimated effects over time. The year 4 effect was statistically significant in this approach.

The finding of effects close to zero in the first year, and generally increasing effects over time, conceivably could result from the continuous improvement approach adopted by the team. As mentioned earlier, some major reforms were implemented over time, in particular having the four schools begin to host the resource teacher on site, and implementing across-school unit planning days, both of which were implemented in year 2. Similarly, the practice of multiple teachers at a school co-designing lesson plans with the resource teacher and co-teaching the lesson multiple times reportedly began in year 3 and expanded in year 4.

How big were the estimated effects? The difference-in-difference approach found a gap in annual changes in Z-scores in favor of the CTO schools. The gains corresponded on average to about 0.11 standard deviation per school year. Given that in 2013 the average Z-score in math in the CTO and comparison schools was -0.32, we can approximate what these estimated gains translate into in terms of percentile rank. Assuming that the student test scores were distributed normally, a gain of 0.11 standard deviation per year, starting from a Z-score of -0.32, would translate into students beginning at percentile 37.4 and rising to percentile 41.6 the next year. These are meaningful gains.

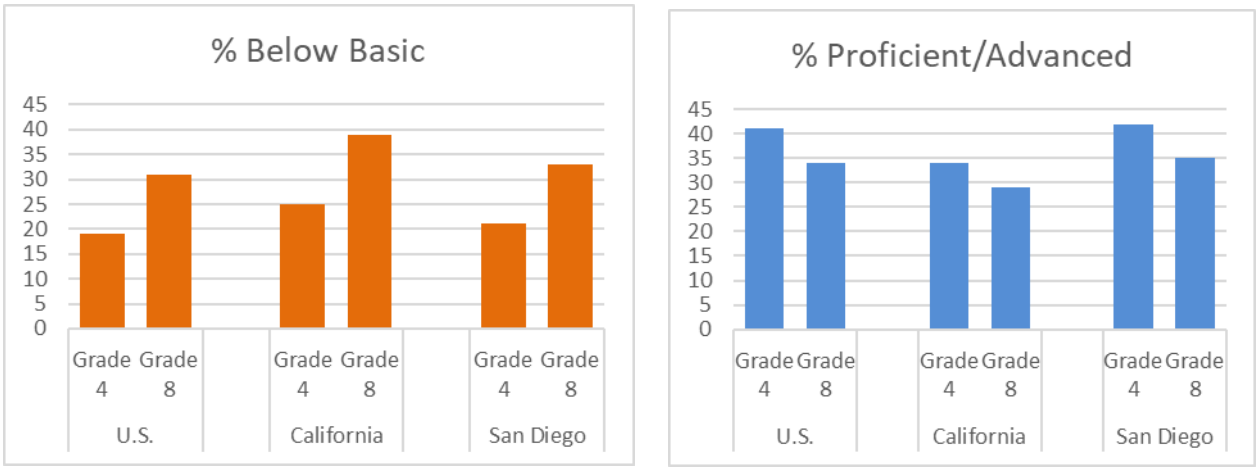
In sum, the difference-in-difference analysis finds a positive influence of the Changing the Odds program on math achievement. A small sample of schools lowers precision of attempts to say much more, but there are suggestive findings indicating that the program became more effective over time. Such a finding, while tentative, meshes with the history of the program itself, which implemented changes over time to increase interaction between the resource teacher and teachers, and to bring teachers together across school to design unit plans.

Our cost analysis suggests that the program costs were equivalent to roughly 4.5% of business-as-usual expenditures on math teachers' wages and benefits. This is a modest expenditure compared to the estimated gains in achievement. It represents about 5% of state supplemental funding given to the district to help meet the needs of underserved students, conservatively assuming these funds were split across district schools on a per student basis.

Replication of this reform in other settings could prove worthwhile. A much larger sample of schools would provide improved insights about which aspects of the program are more and less effective, and whether and how local district context moderates the effects.

Figures and Tables

Figure 1 Percentage of Students "Below Basic" or "At or Above Proficient", 2019 NAEP Mathematics Test, by Grade and Location



Source: National Assessment of Educational Progress (2020).

Figure 2 Scatterplot of Smarter Balanced Grade 6 and 7 Math Scores in 2015 versus 2013 CST Math Scores

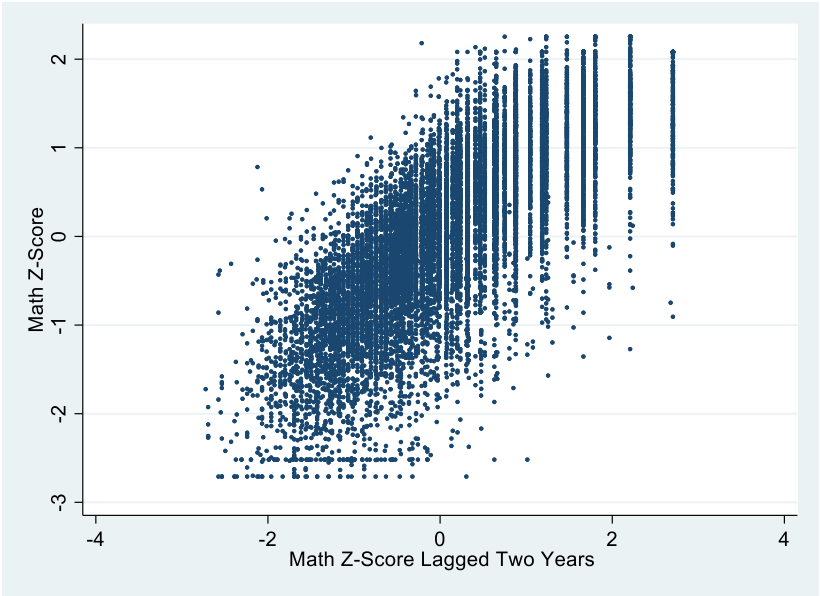


Figure 3 Scatterplot of CST Grade 6 and 7 Math Scores in 2013 versus 2011 CST Math Scores

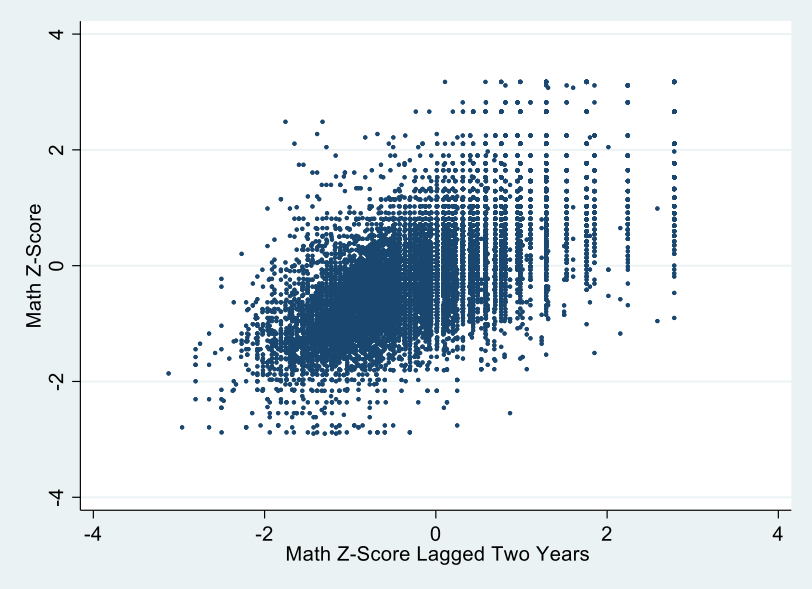


Figure 4 Histogram of 2013 Grade 6 and 7 Math CST Scores in CTO and Comparison Schools

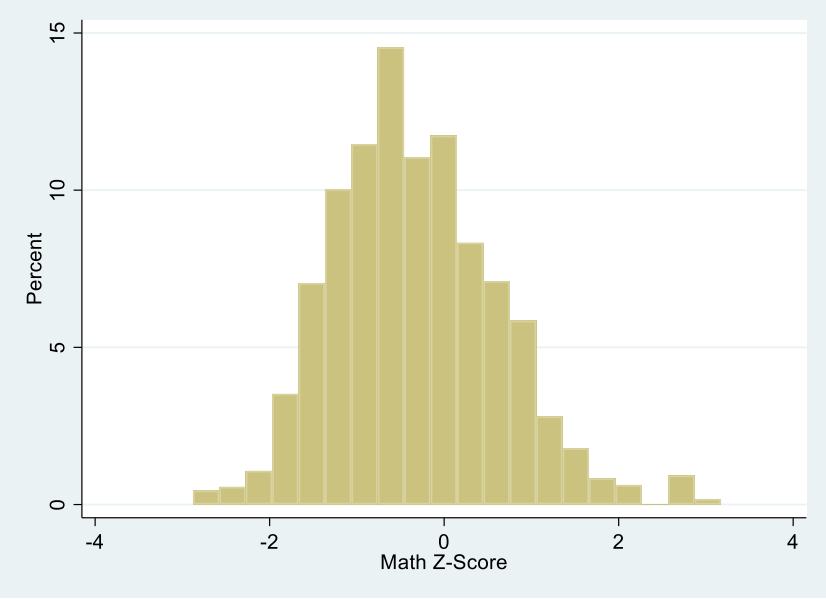
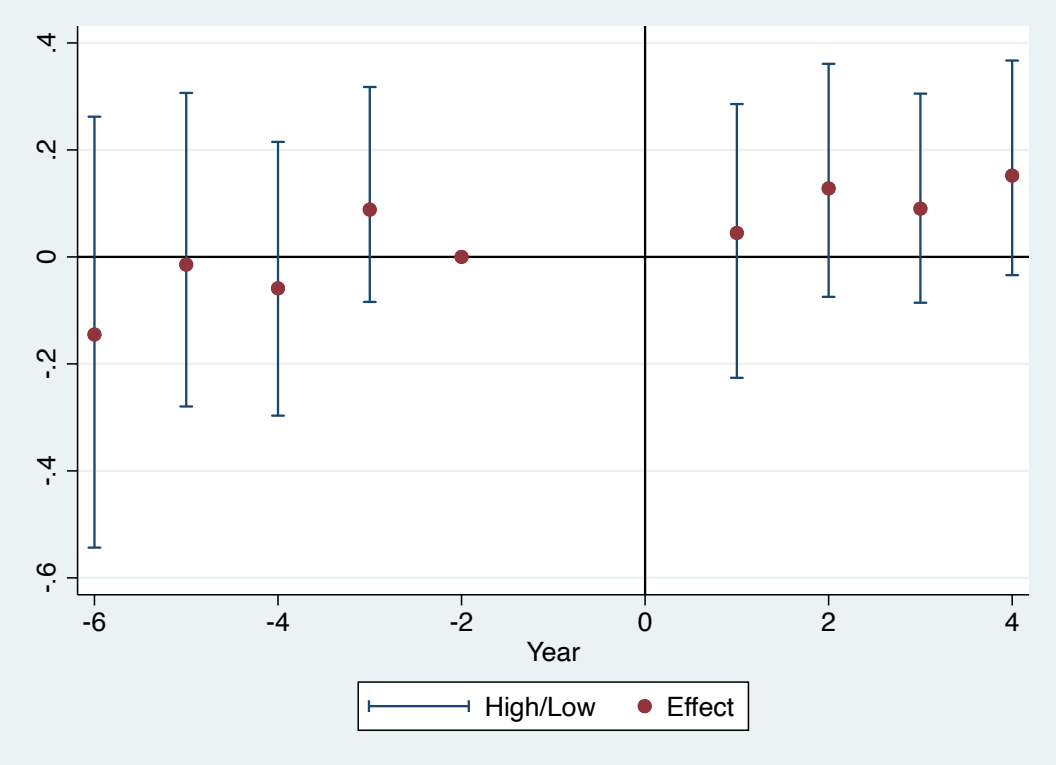
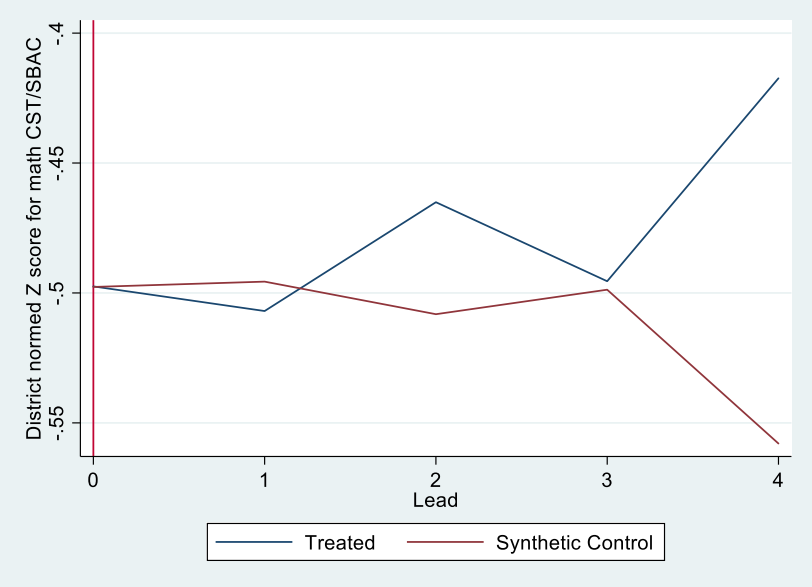


Figure 5 Event Study of Grade 6 and 7 Changes in Students' Math Scores at CTO Schools Relative to Comparison Schools



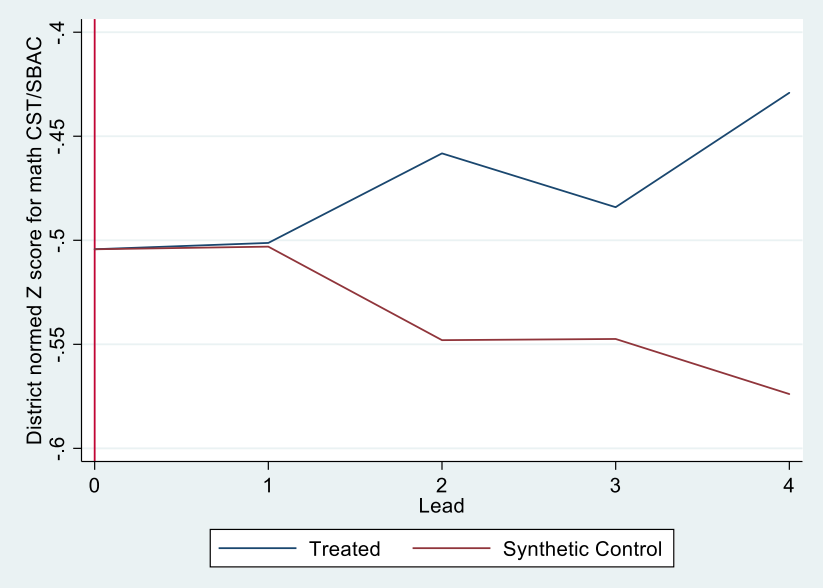
Notes: The dependent variable is changes in individual students' math scores, expressed as changes in Z-scores. Thus the vertical scale shows changes in test scores in terms of the district standard deviation for the given year and grade. The intervention took place in years 1-4, corresponding to 2015-16 through 2018-19. No test score gains are available in years 0 or -1 due to a pause in the state testing system. Changes in test scores in 2012-13 (t=-2) are the baseline against which other years are compared. Hi/lo bars show the 95% confidence interval based on wild bootstrap. Model includes year, grade, gender and race dummies, as well as dummy for CTO schools (any year).

Figure 6 Synthetic Control Model of Impact on the Level of Math Z-Scores in Grades 6 and 7



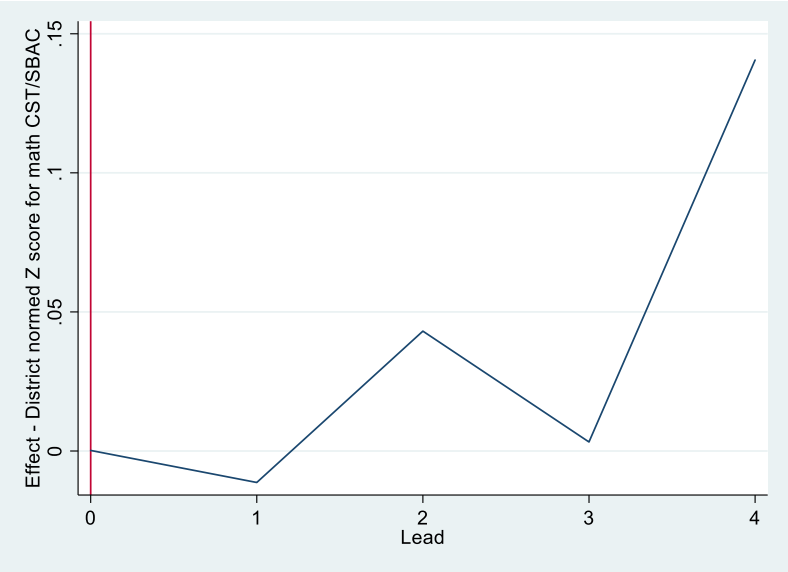
Note: Year 0 corresponds to 2015, while years 1 through 4 refer to the years of the CTO program, 2016 through 2019.

Figure 7 Synthetic Control Model of Impact on the Level of Math Z-Scores in Grades 6, 7 and 8



Note: See note to Figure 6.

Figure 8 Effect Size by Year from Synthetic Control Model of Impact on the Level of Math Z-Scores in Grades 6 and 7



Note: This figure shows the difference between the outcomes in Figure 6.

Table 1 Timeline of the Intervention and State Testing

School Year (Spring)	CTO Program	California Standards Test Grades 2-7 (Grade Specific) and 8-11 (Specific to Math Course)	Smarter Balanced Test Grades 3-8
2009		Yes	
2010		Yes	
2011		Yes	
2012		Yes	
2013		Yes	
2014			
2015			Yes
2016	Yes		Yes
2017	Yes		Yes
2018	Yes		Yes
2019	Yes		Yes

Table 2 Difference-in-Difference Estimates of Changes in Test-Score Growth of Students at CTO Schools

	(1)	(2)	(3)	(4)	(5)	(6)
Coefficient						
CTO*POST	0.126*	0.127*	0.126*	0.105*	0.109*	0.110*
	[0.0376]	[0.0181]	[0.0210]	[0.0350]	[0.0146]	[0.0135]
CTO School	-0.156	-0.146	-0.147			
	[0.0628]	[0.0830]	[0.0799]			
Observations	26,710	26,710	26,710	26,710	26,710	26,710
R-squared	0.015	0.019	0.019	0.029	0.032	0.032
Controls for:						
Year	Yes	Yes	Yes	Yes	Yes	Yes
Grade	Yes	Yes	Yes	Yes	Yes	Yes
Female, Race/Ethnicity	No	Yes	Yes	No	Yes	Yes
Parental Education, Ever EL	No	No	Yes	No	No	Yes
School Fixed Effects	No	No	No	Yes	Yes	Yes
CTO School Fixed Effect	Yes	Yes	Yes	No	No	No

Notes: [ ] entries indicate wild bootstrap p-value
** p<0.01, * p<0.05

Table 3 Event Study of Changes in Test-Score Growth of Students at CTO Schools

Year (Spring)	Year	Effect	Confidence Interval	Effect	Confidence Interval
2009	-6	-0.145	[-.5434, .2621]	-0.133	[-.5273, .2728]
2010	-5	-0.0144	[-.2796, .3065]	-0.00444	[-.2761, .3192]
2011	-4	-0.0587	[-.2967, .215]	-0.0437	[-.2837, .2216]
2012	-3	0.0884	[-.08412, .3176]	0.102	[-.07877, .3349]
2013	-2	0		0	
2014	-1				
2015	0				
2016	1	0.0448	[-.226, .2857]	0.0364	[-.2293, .275]
2017	2	0.128	[-.07466, .3609]	0.121	[-.07812, .3539]
2018	3	0.0903	[-.0857, .3052]	0.0834	[-.08509, .2935]
2019	4	0.152	[-.03412, .3671]	0.144	[-.03534, .3549]
CTO Dummy		Yes		No	
School Fixed Effects		No		Yes	

Note: Other regressors include year dummies, race and gender dummies, and either a CTO dummy or school fixed effects as indicated. ** p<0.01, * p<0.05

Table 4 Results of Synthetic Control Models of the Level of Smarter Balanced Math Scores in 2015-2019, by the Grade Ranges Included

Year (Spring)	Sample: Grades 6 and 7		Sample: Grades 6, 7 and 8	
	Effect	p-value	Effect	p-value
2016	-0.0113	0.9830	0.0018	0.9431
2017	0.0431	0.1638	0.0898 **	0.0027
2018	0.0033	0.9773	0.0633	0.1813
2019	0.1406 **	0.0000	0.1448 **	0.0087

Notes: To guard against placebo effects with relatively poor matches, we limit placebos to those for which the Root Mean Squared Prediction Error is no more than twice that for the given treated school. For the grade 6 and 7 model, inference is based on 4950 placebo averages. For the grade 6 to 8 models, inference is based on 28672 placebo averages.

** p<0.01, * p<0.05

Table 5 Estimated Costs per Year of the CTO Program Overall and Relative to Estimated Costs of the Pre-Existing Teacher Pool

Type of Expenditure	Total Annual Teacher Days Averaged over 4 Years
Resource Teacher at 20% above Cost of Average Teacher	220.8
Summer Math Summit	48.8
Unit Planning	32.3
Co-Teaching	4.3
Desmos Professional Development	1.8
Coach for Resource Teacher	19.4
Books, Teacher Supplies, 1 Day Desmos Workshop	2.2
Average Annual Costs in Teacher Days	329.4
Average Annual Costs in Teacher FTE	1.8
Average Annual Costs in Dollars	$203,897
Average Annual Costs in Teacher FTE per School	0.45
Average Annual Costs in Dollars per School	$50,974
Average Annual Cost as % of Estimated Salaries and Benefits for Pre-Existing Math Teachers	4.5

Notes: The upper panel estimates various elements of the program, averaged over the four years, in teacher FTE equivalents. Many costs were in teacher days, but some costs involved expenditures on other items. We translated back and forth between teacher full time days and dollar costs using the district 184 day teacher contract and the average salary plus benefits (the latter estimated at 60.89%) for a teacher with 10 years of experience with education level Bachelor's+60 or Bachelor's + 54 + Master's. Salary source: San Diego Education Association (2017), p. 116. Benefits rate calculated from district total expenditures on salaries and on benefits for 2018-19, from San Diego Unified School District (2019), p. 18.

It is possible that some district expenditures on benefits accrue to former district employees, which would mean that the benefits rate is overstated. But this would only serve to overstate our estimated dollar costs of the teacher time allocated to the CTO program, making the program look less cost effective than it was.

Appendix Table 1 Comparison of 2012-13 Characteristics of the Four Treated Schools and the Four Comparison Schools, Along with p-Values for Test of Differences

Variable	Treatment	Control	P-Value for Test Treatment=Control
API Achievement Measure, 5Spring 2013	730	735	0.97
% Basic or Better on State Math Test Spring 2013	65	70	0.81
School Characteristics Index Spring 2013	165	165	0.66
% EL and Recently Reclassified	65	60	0.64
% Hispanic	80	65	0.13
% African American	10	10	0.58
% White	5	10	0.56
% Asian	5	10	0.48
% Eligible for Meal Assistance	100.0	90	0.36

Notes: Treatment and control means are rounded to the nearest multiple of 5 to reduce the chances of inadvertently revealing the identity of the schools. For accountability purposes, California combines current English Learners and Reclassified students who have not yet tested proficient or better on the ELA test for three years running.

API refers to the Academic Performance Index, which is a function of test scores in math and English Language Arts tests that were given in grades 2-11 in California schools between 2002 and 2013.

References

Abadie, Alberto, Diamond, Alexis & Hainmueller, Jens (2010). "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, (105:490), pp. 493-505.

Abadie, Alberto and Gardeazabal, Javier. (2003). "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review* , March (93:1), pp. 113-132.

Ashlock, Robert, (2010), **Error Patterns in Computation: Using Error Patterns to Help Each Student Learn (10th Edition)**, Boston: Allyn and Bacon.

Betts, Julian R., Hahn, Youjin, and Zau, Andrew C. (2017). "Can Testing Improve Student Learning? An Evaluation of the Mathematics Diagnostic Testing Project". *Journal of Urban Economics, 100*, 54-64.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education*, *85*(2), 205-225.

Boaler, J. (2015). Mathematical mindsets: Unleashing students' potential through creative math, inspiring messages, and innovative teaching. Jossey Bass: San Francisco.

Bryk, Anthony S., Louis M. Gomez and Alicia Grunow (2011), "Getting Ideas into Action: Building Networked Improvement Communities in Education," in Maureen Hallinan (Ed.), **Frontiers in Sociology of Education**, Springer Publishing.

Cameron, A. Coin, Jonah B. Gelbach and Douglas L. Miller (2008), "Bootstrap-Based Improvements for Inference with Clustered Errors", *Review of Economics and Statistics*, Vol. 90, pp. 414-427.

Cavallo, Eduardo, Galiani, Sebastian and Noy, Ilian (2013) "Catastrophic Natural Disasters and Economic Growth," *Review of Economics and Statistics,* (95:5), pp. 1549-1561.

Cosner, S. (2011). Teacher learning, instructional considerations and principal communication: Lessons from a longitudinal study of collaborative data use by teachers. *Educational Management Administration & Leadership*, *39*(5), 568-589.

Daly, A. J. (2012). Data, dyads, and dynamics: Exploring data use and social networks in educational improvement. *Teachers College Record, 114*(11), 110305.

Datnow, A., & Park, V. (2014). *Data driven leadership*. San Francisco, CA: Jossey Bass.

Datnow, A., & Park, V. (2014). *Professional collaboration with purpose: Teacher learning towards equitable and excellent schools.* New York: Routledge.

Deming, W.E. (1986). **Out of the Crisis**, MIT Center for Advanced Engineering Study.

Feldman, J. & Tung, R. (2001). *Whole school reform: How schools use the data-based inquiry and decision making process.* Paper presented at the 82nd annual meeting of the American Educational Research Association in Seattle, WA.

Honig, M. I., & Venkateswaran, N. (2012). School–central office relationships in evidence use: Understanding evidence use as a systems problem, *American Journal of Education*, *118*(2), 199-222.

Light, D., Honey, M., Heinze, J., Brunner, C., Wexler, D., Mandinach, E., & Fasca, C. (2005). *Linking data and learning: The Grow Network study*. New York: Educational Development Center, Inc.

Marsh, J.A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, *114*(11), 1-48.

Means, B., Padilla, C. and Gallagher, L. (2010), *Use of Education Data at the Local Level: From Accountability to Instructional Improvement*, U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, Washington, DC.

National Assessment of Educational Progress (NAEP),(2020) **Mathematics and Reading Assessment**s. U.S. Department of Education, Institute of Education Sciences. Results downloaded from http://https://www.nationsreportcard.gov/ndecore/xplore/NDE.

OECD (2016), **PISA 2015 Results (Volume I): Excellence and Equity in Education**, PISA, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264266490-en.

Provasnik, S., Malley, L., Stephens, M., Landeros, K., Perkins, R., and Tang, J.H. (2016). **Highlights From TIMSS and TIMSS Advanced 2015: Mathematics and Science Achievement of U.S. Students in Grades 4 and 8 and in Advanced Courses at the End of High School in an International Context (NCES 2017-002).** U.S. Department of Education, National Center for Education Statistics. Washington, DC. Retrieved September 2020 from http://nces.ed.gov/pubsearch.

Roodman, David, MacKinnon, James G., Nielsen, Morten Ørregaard, and Webb, Matthew D. (2019), "Fast and Wild: Bootstrap Inference in Stata Using Boottest," *Stata Journal* (19:1), pp. 4–60 DOI: 10.1177/1536867X19830877.

Rose, Heather, and Julian R. Betts, (2004). "The Effect of High School Courses on Earnings," *Review of Economics and Statistics*, 86(2), 497-513.

San Diego Education Association (2017). "Collective Bargaining Agreement between the Board of Education San Diego Unified School District and the San Diego Education Association," downloaded 5/17/2021 from http://www.sdea.net/wp-

content/uploads/SDEA-Complete-Contract-2017-2020-with-Contract-Addendum-with-2019-Reopeners.pdf.

San Diego Unified School District (2019). "San Diego Unified School District Financial Statements, June 30, 2019 ", downloaded 5/17/21 from https://sandiegounified.org/UserFiles/Servers/Server_27732394/File/Departments/Controller/SDUSD-6.30.2019-FS-Final.pdf.

Webb, M. D. (2014). "Reworking Wild Bootstrap Based Inference for Clustered Errors," Queen's University, Department of Economics, Working Paper No. 1315. https://ideas.repec.org/p/qed/wpaper/1315.html.

White, P., & Anderson, J. (2011). Teachers' use of national test data to focus numeracy instruction. In J. Clark, B. Kissane, J. Mousley, T. Spencer & S. Thornton (Eds.), *Mathematics: Traditions and [new] practices* (pp. 777-785). Adelaide: AAMT and MERGA.