CPRE
CONSORTIUM FOR POLICY RESEARCH IN EDUCATION
TEACHERS COLLEGE
COLUMBIA UNIVERSITY

FINAL IMPACT RESULTS FROM THE i3 IMPLEMENTATION OF *TEACH TO ONE: MATH*

JANUARY 2019

Douglas D. Ready
Katharine Conn
Shani S. Bretas
Iris Daruwala

©2019 Consortium for Policy Research in Education at Teachers College, Columbia University. All errors of either fact or interpretation are solely those of the authors and not those of New Classrooms, the Elizabeth Public Schools, the Institute of Education Sciences, or the U.S. Department of Education.

For further information about this paper or the larger study of which it is a part, contact the first author at ddr2111@tc.columbia.edu or Box 124, 525 W. 120th Street, New York, NY 10027

ACKNOWLEDGMENTS

This study was made possible by the assistance of many individuals. First and foremost, we thank the students, teachers, principals and central office administrators in Elizabeth, New Jersey who so warmly opened their classrooms and schools to us. We also owe a tremendous debt to Amy Gil and Venkatesh Sagaram at the Elizabeth Public Schools Office of Research, Evaluation, and Assessment who provided the necessary data and who spent four years answering our many questions. At New Classrooms, Jennifer Stillman afforded invaluable support throughout the entire project. At CPRE, Anna Moyer provided crucial research and editing assistance.

ABOUT THE CONSORTIUM FOR POLICY RESEARCH IN EDUCATION (CPRE)

The Consortium for Policy Research in Education (CPRE) brings together education experts from renowned research institutions, including Teachers College, Columbia University; the University of Pennsylvania; Harvard University; Stanford University; the University of Michigan; the University of Wisconsin-Madison; and Northwestern University. CPRE was launched in 1985 as the first national federally funded R&D center for state and local education policy. Since then, CPRE has studied the design, implementation, and effects of hundreds of policies and programs. CPRE researchers and staff at Teachers College (CPRE-TC) are continuing this long tradition by conducting rigorous research and evaluation that aims to improve elementary and secondary education through increased educational effectiveness, equity, and access.

To learn more about CPRE-TC, visit our website: www.tc.columbia.edu/cpre

TABLE OF CONTENTS

EXECUTIVE SUMMARY

New Classrooms Innovation Partners was awarded an Investing in Innovation (i3) grant from the U.S. Department of Education to implement its *Teach to One: Math* (TtO) model in five schools in Elizabeth, NJ for three academic years, beginning in September 2015 and continuing through June 2018. The Consortium for Policy Research in Education (CPRE) at Teachers College, Columbia University, recently concluded a four-year evaluation of these efforts. The evaluation had two primary strands: 1) a methodologically rigorous quantitative study of TtO's impact on student mathematics performance; and 2) a systematic qualitative analysis of TtO's implementation processes, combined with yearly New Classrooms and school staff interviews, student focus groups, and classroom observations in each of the five TtO schools. This report describes the results of the first strand, focusing exclusively on the quantitative impact results.

Data and Methods

Our analyses employed a sample of 36,158 student-level measurements, nested within 209 school-by-year cohorts, located within five TtO schools and a demographically and academically comparable group of 16 non-TtO schools, all in Elizabeth, NJ. Although a small number of fifth graders participated in TtO, our analyses focus solely on sixth, seventh, and eighth graders in these K-8 schools. The data include student demographic measures and test scores for seven years prior to the i3 implementation and three years of data from the implementation period. Our primary analyses entailed a comparative interrupted time series (CITS) approach (also known as the difference-in-differences technique). We used these models to compare changes in state-mandated standardized test scores among TtO schools before and after program implementation to the changes in outcomes among a similar group of schools in Elizabeth that did not implement TtO during the same period.

Findings

Performance patterns across the five TtO treatment schools were quite inconsistent, and the TtO estimates for each of the three implementation years were statistically non-significant. We found similar results using a measure indicating the combined (averaged) three-year impact and models that excluded a number of relatively higher-achieving control schools. Importantly, however, during the implementation period, participating TtO schools in Elizabeth asked for a variety of program configurations and alterations. Specifically, different schools requested adjustments that either emphasized or deemphasized the extent to which their students received grade-level mathematics content. This heterogeneity in both outcomes and the specific nature of the treatment itself leaves us unable to draw generalizable conclusions, positive or negative, regarding TtO's impact on student mathematics performance as measured by state-mandated grade-level assessments.

Limitations

These results come from a single school district, with a particular set of teachers in a single three-year implementation period. TtO may well have had different results—either positive or negative—in a different context at a different point in time. And, although the models accessed over 36,000 student-level test scores and seven years of prior achievement data, the analyses still include only five treatment and 16 control schools. While adequate from an analytical perspective, our ability to identify statistically significant treatment effects that were modest in size—again, either positive or negative—was somewhat limited by the school sample size.

Readers should bear in mind an additional factor when interpreting these results, or indeed the results of any study of personalized learning that employs standardized test scores as outcomes. A central characteristic of TtO is that students who are missing foundational math skills are provided the time and space to experience below grade-level content. One result of this differentiation is that TtO and non-TtO students with the same initial test scores likely receive quite different mathematics content during the academic year. Because TtO students may receive content that is below grade-level, their performance on state-mandated, grade-level standardized assessments may lag behind their non-TtO peers who are being exposed to new, grade-level skills over the course of the academic year. This would be the case regardless of whether non-TtO students possess the fundamental skills needed for long-term success in mathematics. In other words, mere exposure to increased amounts of grade-level content—although possibly inappropriately advanced—may provide a near-term impact on grade-level assessments. But it may not be the best approach for all students in the longer term.

A final consideration is that the TtO model continues to evolve. Its design elements and algorithms are not fixed and can presumably improve over time. As such, it is unclear the extent to which the model examined during this implementation period in Elizabeth, NJ is representative of the TtO model currently being used in schools today. In sum, these analytic complications, which would likely be common to many evaluations of personalized learning models, obviously warrant further consideration.

Conclusions

TtO is a blended learning intervention that represents a fundamental disruption of traditional classroom processes. Indeed, it is a unique and distinct approach to mathematics teaching and learning. Unlike so many other educational reforms, TtO is a deep reform that seeks to directly influence the instructional core and to improve how students and teachers interact around academic content. Moreover, its focus is middle school mathematics, which has received far less research attention, particularly compared to the elementary grades. Future research should continue to explore TtO's promise, and seek to establish its effectiveness in other contexts and with different implementation and evaluation strategies.

FINAL IMPACT RESULTS FROM THE I3 IMPLEMENTATION OF *TEACH TO ONE: MATH*

BACKGROUND

Practitioners and researchers have long explored the appropriate way to manage academic diversity in the classroom (Corno, 2008). Because students in the U.S. are typically grouped into same-age classrooms regardless of academic background, teachers confront a tremendous amount of variability in student skills. Indeed, despite considerable stratification and segregation across schools and classrooms, nationally representative data suggest that almost 62% of variability in fifth-grade mathematics ability remains within classrooms, with the additional variability split relatively evenly between classrooms in the same school and between schools (Martinez, Stecher, & Borko, 2009). Such academic diversity has historically been handled via ability grouping in the lower grades (Barr & Dreeben, 1983; Pallas, Entwisle, Alexander, & Stluka, 1994) and curricular differentiation and "tracking" in the upper grades (Lee & Ready, 2009; Oakes, 1985). Unlike these traditional approaches, which typically do not serve low-achieving students well, advocates have instead argued for "adaptive instruction" (Snow, 1980) and "personalized learning" approaches (Gates Foundation, 2014), which seek to respond to individual students' needs amidst the diversity of the collective social classroom environment (Corno, 2008). Similar to Vygotsky's (1978) "zone of proximal development," personalized approaches aim to situate students in cognitive spaces that are slightly beyond their current skills, and "scaffold" their learning by concentrating efforts into more manageable sets of tasks (Bruner, 1978).

Although various efforts to personalize learning have existed for decades, technological innovations over the past several years have made it much easier to individualize instruction based on each student's level of content mastery and developmental trajectory (Wolf, 2010). These contemporary approaches entail individualized and personalized learning plans based on student-level data, recognize progress that is based on demonstrated knowledge rather than seat time, and employ multiple and flexible pedagogical and learning environments (Gates Foundation, 2014). With the assistance of "blended learning" approaches, which combine computer-based and live teacher-directed instruction, instructional models that incorporate personalized learning approaches have been implemented in hundreds of schools across the U.S.

Teach to One: Math

One such program that involves both personalization and blended learning is *Teach to One: Math* (TtO), developed by New Classrooms Innovation Partners. TtO currently serves students in 39 charter and traditional public schools across 11 states. TtO has been highlighted in the New York Times, Wall Street Journal, the Atlantic, and Time Magazine, which named an earlier version of the program as one of the top 50 "inventions of 2009." At a recent conference, Bill Gates referred to TtO as "the future of math" (Newcomb, 2016). TtO is often cited by technology advocates and reformers as one of a small number of archetypal examples of technology-based personalized learning. TtO redesigns classroom instruction in an attempt to match students with the specific content that best supports their academic growth. The key design features of the program have remained largely consistent since its inception in 2009 as a summer pilot program called "School of One" within the New York City Department of Education. Central to the program is the "skills map," which outlines the roughly 300 discrete skills that students must master and the dependencies among them. A second component is each student's individual learning profile, which is initially generated from a baseline assessment administered at the start of the year, then updated frequently (often daily) based on student performance.

TtO physically reorganizes the learning environment into one large room containing multiple teachers and students from different classes simultaneously, at times including students from multiple grade levels. Upon entering the room, students meet in their Math Advisory (MA) sections, open personal laptop computers, log into the TtO online portal, and consult their personal "playlists," which tell them what content they will be learning that day and how they will be learning it. The program incorporates a variety of learning modalities, including independent work, peer-to-peer learning, and traditional teacher-led instruction. Large TV screens around the classroom direct students to a designated area designed for their first assigned modality. Although TtO provides teachers with instructional resources to use in teacher-led modalities, teachers may customize them or use different materials of their own design if they so choose. The TtO period is split into two, approximately thirty-minute sessions. After the first session, students again consult the large TV screens and report to a different area of the room for their second modality and skill assignment for the period. At the end of each day, students return to their MA sections to take a short, multiple choice "exit ticket" to determine their mastery of that day's content. The program then uses the exit ticket results to update students' individual learner profiles and determine their assignments for the next day.

Prior Studies

TtO has not previously undergone a rigorous impact evaluation. However, several studies evaluated the earlier version of TtO, known then as School of One. The NYC DOE received an i3 grant to expand School of One into four New York City middle schools during the 2012-13 and 2013-14 school years. Evaluators conducted a cluster randomized controlled trial with four treatment and four matched control schools. Findings from the study indicated no statistically significant impacts of the program on either New York State standardized assessments or the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessment (Rockoff, 2015). However, the study encountered considerable implementation problems, such as the loss of several months of participation due to severe Hurricane Sandy damage and a non-compliant treatment school. As the study's author suggested, these null findings should, therefore, be interpreted with caution. An even earlier evaluation involved three New York City School of One pilot schools and a comparative interrupted time series (CITS) design during a single implementation year (2010-11). The study authors reported considerable outcome heterogeneity across schools, concluding that it was "impossible to draw definitive conclusions about the overall effectiveness of the program or the conditions under which it might be more effective" (p. ES-3; Cole, Kemple, & Segeritz, 2012). It is important to bear in mind that this study explored a much earlier version of TtO. More recently, a descriptive report using data from all seven TtO schools operating in 2012-13 and all fifteen operating in 2013-14 found that students attending TtO schools exhibited gains that surpassed national NWEA MAP norms (Ready, 2014). However, as the author noted, such an approach clearly cannot establish the causal impact of TtO on student mathematics performance, given the unknown comparability of comparison students and schools that constituted the national sample.

The TtO Model in Elizabeth, NJ

Even with the exposure to grade-level skills that is built into the TtO algorithm, many districts and schools request additional adjustments to either emphasize or deemphasize grade level content. In the first two years of this study, nationally across all TtO schools, these adjustments came in the form of "floors" which limited the content that students had access to. For example, students needing skills that were two years below their effective grade level could access that skill only if the floor was set at two or more years below grade level. Similarly, some schools asked that TtO incorporate "ceilings" that limited how far ahead students could go in relation to grade level standards. While some students were fully capable of accelerating into above-grade content, particular schools and districts preferred that such students instead review and deepen their understanding of grade level material. Finally, in addition to these changes, some schools requested a more focused "test-prep" period during the weeks directly prior to the administration of high-stakes state assessments. This further prioritized grade level material above and beyond the floors and ceilings that may have been imposed throughout the school year.

In Elizabeth, over the first two years of implementation, the incorporation of floors and ceilings based on school and district preferences varied widely across schools, semesters, and even across grade level cohorts within schools. However, an overall floor was imposed at all Elizabeth schools at three years below grade level (although many students began the school year below that level). In addition, during the first year of implementation two out of the five TtO treatment schools requested a test-prep period of five to six weeks where students worked exclusively on grade-level content. In the second year, four of the five schools opted for this test-prep period.

In the third and final year of implementation, and in part based on the learnings from Elizabeth, New Classrooms provided more structure to the disparate set of floors and ceilings that schools and districts requested, and created a set of "Outcome Strategies" to pilot in select schools. One option mirrored the traditional test-prep window many schools previously employed: students were allowed to work up to three years below their registered grade for the majority of the year, but were abruptly moved to grade-level skills for the last four to six weeks before the state assessment administration. In addition to this approach, New Classrooms developed two other Outcome Strategies. In one approach, grade level floors were gradually introduced over the school year to avoid an abrupt shift to grade level skills at the end of the year. In another, students began each unit with grade level skills and were only moved to specific below-grade content if they were missing the foundational skills necessary to grasp specific grade-level content. Two Elizabeth schools employed this model during the third year of implementation.

Lastly, it is important to note that in New Jersey, districts have the option of focusing the eighth-grade and end-of-year testing on either the eighth-grade state standards or Algebra (which includes more advanced mathematical concepts than the eighth-grade standards). Elizabeth chose the latter for its eighth graders, and thus for purposes of TtO, "grade-level skills" were defined as those skills that relate to Algebra. This choice may be appropriate for those with the requisite predecessor knowlege. But for other students, skipping over the eighth-grade curriclum may have further exacerbated the gap between their zone of proximal development and the content they received in Teach to One.

DATA AND METHODS

The Implementation Context

These analyses employed student-level data from five treatment and 16 comparison schools in Elizabeth, New Jersey, a high-minority enrollment, high-poverty school district within the New York City metropolitan area. In Elizabeth, 26 K-8 schools and six high schools serve approximately 23,000 students, 90% of whom are either black or Hispanic, with a similar proportion eligible for free or reduced-price lunch. Over 13% of students are enrolled in bilingual classes. New

Classrooms worked with the district to identify a subset of treatment schools that met two criteria. First, the schools were willing to participate in the implementation for the full three-year study period. Second, the schools were required to have the physical/structural capacity to combine several classrooms into one large TtO space, or were willing to devote an existing large space to TtO during TtO instructional time (as one school did with its media center). As we describe below, the process produced a set of treatment and control schools there were quite well matched in terms of student academic and socio-demographic backgrounds. Additionally, given that TtO is a "deep" intervention, with a substantial investment in infrastructure and technological components, there was no cross-school contamination of the TtO model to other non-TtO schools. In other words, replication at other sites within EPS was not possible without the support of New Classrooms.

The fact that both TtO and comparison schools are drawn from the same relatively homogeneous district supports our ability to make claims about TtO's impact on student achievement. To further ensure comparability, we eliminated two district-wide gifted and talented schools and three schools that were using TtO but were not involved in the i3 implementation. The resulting sample of five TtO and 16 control schools was quite balanced in terms of student characteristics in the years prior to TtO implementation. However, to test the robustness of the impact estimates, we also constructed a series of models that eliminated six higher-performing control schools. We report the results of all models from both samples below.

The five treatment schools implemented TtO in grades 6-8 for all three implementation years. During the first two years, three TtO schools also had fifth graders participate; however, during the third implementation year, one of these three schools dropped fifth grade from the implementation. Given the unequal treatment of fifth grade during the study period (and the attendant reduction in fifth-grade sample sizes), we eliminated fifth grade students in both the treatment and control schools from our analyses and focused exclusively on grades 6-8.

Comparative Interrupted Time Series Models

To provide evidence of the extent to which TtO improves student learning in mathematics, we used a comparative interrupted time series (CITS) approach (Bloom, 2003; Shadish, Cook, & Campbell, 2002), with state-mandated standardized assessment scores as outcomes. These models were constructed within a multi-level framework (see Raudenbush & Bryk, 2002) with students nested within school cohorts, which were nested within schools. Within the context of the current evaluation, the CITS approach compared changes in outcomes among TtO schools before and after program implementation to the changes in outcomes among a similar group of schools that did not implement TtO in Elizabeth during the same period. We employed a multi-level baseline means model, which reduces sensitivity to noise in student achievement and reduces the risk of misspecified slopes (see Bloom, 2003).

The model can be described as:

Level-1 (students): $Y_{ijk} = \pi_{0jk} + \pi_{.jk}(X_{ij}) + e_{ijk}$

Level-2 (cohorts): $\pi_{0jk} = \beta_{00k} + \beta_{01k}(TrtYr1_{jk}) + \beta_{02k}(TrtYr2_{jk}) + \beta_{03k}(TrtYr3_{jk}) + \beta_{0.k}(X_{jk}) + r_{0jk}$

Level-3 (schools): $\beta_{00k} = \gamma_{000} + \gamma_{001}(TtO_k) + u_{00k}$

$\beta_{01k} = \gamma_{010} + \gamma_{011}(TtO_k)$

$\beta_{02k} = \gamma_{020} + \gamma_{021}(TtO_k)$

$\beta_{03k} = \gamma_{030} + \gamma_{031}(TtO_k)$

where: Y_{ijk} = test score for student *i* in cohort *j* in school *k*

X_{ij} = a vector of student covariates, centered around their within-cohort means

e_{ijk} = the error term associated with child *ijk*, assumed to be normally distributed with a mean of zero and a constant Level-1 variance, σ^2

$TrtYr1_{jk}$ = 1 if treatment year 1 (2015-2016 academic year), 0 if pre-treatment year (2008/09-2014/15 academic years)

$TrtYr2_{jk}$ = 1 if treatment year 2 (2016-2017 academic year), 0 if pre-treatment year (2008/09-2014/15 academic years)

$TrtYr3_{jk}$ = 1 if treatment year 3 (2017-2018 academic year), 0 if pre-treatment year (2008/09-2014/15 academic years)

X_{jk} = a vector of cohort (within-school) covariates (aggregates of student-level measures), centered around their within-school means

r_{0jk} = error associated with cohort j in school k

γ_{000} = the mean for comparison (non-TtO) schools during pretreatment years

γ_{001} = the average difference between TtO and non-TtO schools for pretreatment years

u_{00k} = error associated with school *k*

γ_{010} = the average difference between pretreatment and year 1 for the comparison schools

γ_{011}, γ_{021}, and γ_{031} = ***our focus: the separate treatment effects for year 1, 2 and 3 (the difference-in-differences estimators).*** We also ran a series of models with a single treatment indicator (i.e., the combined TtO effect in years one through three).

The CITS models leveraged seven years of prior student-level achievement and socio-demographic data. As in many states, New Jersey's school accountability system has experienced considerable flux over the past several years, with changing standardized assessments over time. The models presented here employed data from two large-scale mathematics assessments. For pre-test academic years 2008-09 through 2013-14, we used student-level results on the New

Jersey Assessment of Skills and Knowledge (NJASK).[1] Beginning with the 2014-15 academic year, New Jersey adopted assessments organized by the Partnership for Assessment of Readiness for College and Career (PARCC).[2] For pre-treatment year 2014-15, as well as all three treatment years (2015-16, 2016-17, and 2017-18), we used the student-level PARCC scale scores as outcomes. All test scores were standardized (z-scored) within year and grade. Data were received directly from EPS.

Covariates

Student-level covariates for the CITS models included gender (female=1, male=0), limited English proficiency (LEP) special education status (yes=1, no=0), and separate indicators of race/ethnicity (black, Hispanic and Asian/Pacific Islander students, all compared to white students). We also incorporated separate measures of free- and reduced-price lunch status (yes=1, no=0), mobility (student changed schools at any point during grades 5-8; yes=1, no=0), and grade level (sixth and eighth, compared to seventh). The cohort-level (Level-2) models included aggregate indicators of these student-level measures (except grade) to capture changes in student social and academic background characteristics over time within schools. At the school-level (Level-3), the sole indicator is whether the school participated in TtO. The student-level and cohort-level measures were group-mean centered within the CITS models (analogous to fixed effects). There was virtually no missing covariate data. Students missing test score outcomes were eliminated from the sample. [3]

RESULTS

Table 1 provides baseline descriptive information on student socio-demographic and academic characteristics across TtO and non-TtO comparison schools prior to the first implementation year (pre 2015-16). Although all models employed covariate adjustments, it is important that the treatment and control schools be as similar as possible on all pre-treatment indicators. This is indeed what Table 1 suggests, with no substantively or statistically significant differences in student background characteristics between students enrolled in TtO and non-TtO schools. All schools had somewhat larger proportions of male compared to female students, and LEP and special education enrollments were roughly 8-12% of school enrollments. These schools also served high-poverty student clientele, with roughly three out of four students eligible for free- or reduced-price lunch, and a largely non-white student demographic, with black and Hispanic students constituting roughly 90% of school enrollments.

[1] For more information on NJASK, see www.nj.gov/education/assessment/es/njask/
[2] For more information on PARCC, see www.parcconline.org/about
[3] Fewer than 9% of cases were missing test scores, including students not enrolled in an EPS school during the testing period.

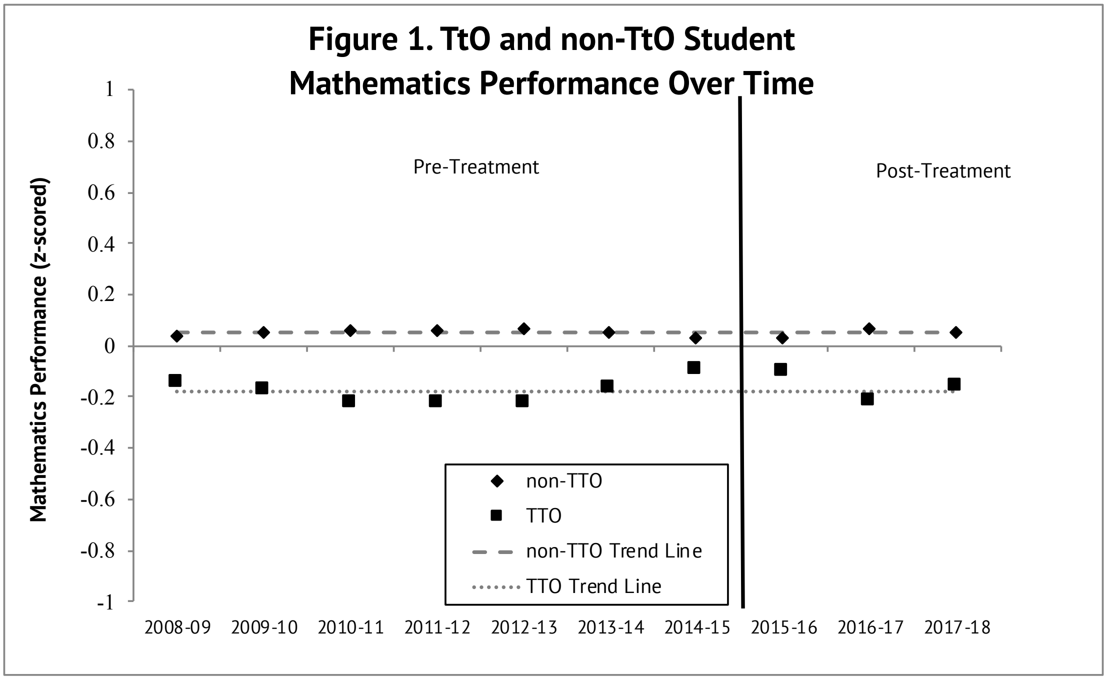
To establish baseline equivalence in prior mathematics achievement, we constructed a reduced (unadjusted) form of the multilevel CITS model described above that employed only pre-implementation years as outcomes and the TtO school-level indicator as the sole predictor. The estimates from these models, which are also presented in Table 2, indicate no statistically significant difference in student mathematics achievement between TtO and non-TtO schools. We also constructed a fully adjusted version of this baseline model, which produced virtually identical estimates. This is unsurprising, given the good pre-treatment covariate balance across TtO and non-TtO schools.

Table 1. Baseline (Pre-Treatment) Academic and Socio-Demographic Characteristics for TtO and non-TtO Students

	TtO Schools (*n*=5)	Non-TtO Schools (*n*=16)
Baseline Measurements	(*n*=5,297)	(*n*=19,379)
% Female	47.3	48.7
% Limited English Prof.	11.6	8.9
% Special Education	11.4	9.2
% Free Lunch	73.6	77.9
% Reduced-Price Lunch	9.7	10.3
% Mobile	19.5	19.0
% Asian	2.5	1.3
% Black	22.6	22.6
% Hispanic	64.4	68.6
Baseline Math Ach. (z-scored)	-0.17	0.04

No differences significant at the $p<.05$ level. Full sample (including implementation years) includes 36,158 student-level measurements nested within 209 school-by-year cohorts, within five TtO and 16 non-TtO schools.

Figure 1 below displays standardized test score means for students in both TtO and non-TtO schools. In the years prior to TtO implementation, student mathematics achievement in TtO and non-TtO schools was quite constant, fluctuating only slightly within a very narrow range. The somewhat higher rate of variability among TtO schools across years is explained largely by the smaller number of treatment compared to control schools. The dotted lines in each figure represent the mean performance of TtO and non-TtO schools prior to TtO implementation projected into the implementation years. Note that TtO school performance in the first implementation year was (at least visually) somewhat higher than what was predicted based on the trend-line established using TtO school prior performance; in year two it was slightly below; and in year three, TtO school performance was roughly where it was pre-implementation. The CITS models compare differences in scores among TtO schools pre- and post-implementation, to the differences in scores among non-TtO schools, pre- and post-implementation.

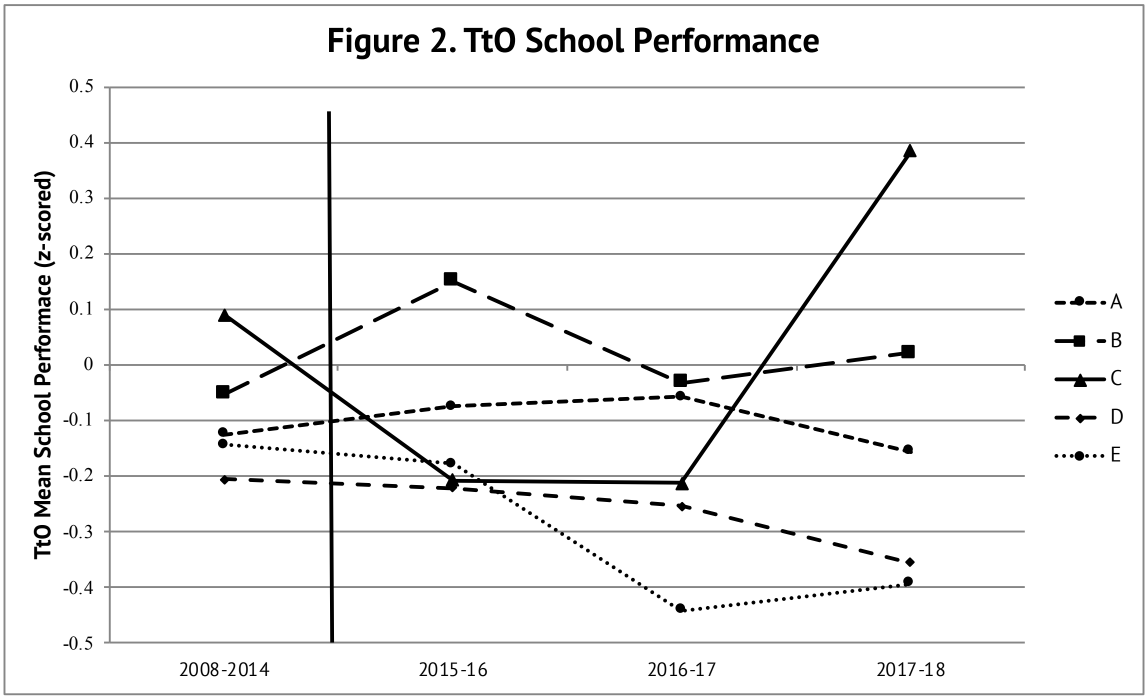


Treatment Group Heterogeneity

Achievement patterns differed both within and between the five TtO schools during the implementation period. Figure 2 displays school means for pre-treatment years on the far left of the figure, as well as mean mathematics achievement among TtO schools for each treatment year.[4] Note that during the first implementation year (2015-16), performance for schools A, D, and E was quite close to average pre-treatment levels. However, performance among students in school C was below pre-treatment levels, while test scores in School B were above pre-treatment levels. After year two, achievement at Schools A, C, and D was stable from the prior year, while achievement at Schools B and E declined. With the final implementation year, School B and E recovered somewhat from the prior year decline, while Schools A and D experienced minor declines. But the anomaly here is obviously School C, whose mean achievement increased by roughly 0.6 standard deviations during the third implementation year. The explanation for a portion of this dramatic increase is relatively straightforward. For several years School C had served as (what the district termed) a "port of entry" school, meaning it enrolled substantial numbers of recent immigrant students, many of whom had limited English skills, as well as substantial numbers of students with limited histories of formal schooling. Due to this role, during the first two implementation years School C had LEP enrollments that averaged almost 40%—three to four times the district school average—and free-lunch rates over 50%. In the third implementation year, however, the school no longer served in this capacity. As a result, the

[4] The patterns are very similar whether we use the average of pre-implementation years for each school for the initial baseline markers on the far left (as we did here) or only data from the 2014-15 school year.

school's LEP enrollment declined precipitously to only 4.4%, and fewer than 30% of students were free-lunch eligible. As we describe below, we ran a series of alternative models to determine the extent to which this school influenced our results.



We also explored whether the slightly different permutations of the TtO model implemented in Elizabeth explained these differential outcomes among the five TtO schools. Table 2 below summarizes the different approaches schools used in terms of grade-level content (as described above). When we map the strategies in Table 2 onto the achievement patterns in Figure 2, we find no relationship—variability across schools in the content delivery strategy was unrelated to these achievement patterns. For example, of the three schools that did not have a test-prep period during the first implementation year, performance at one school rose slightly (School A), declined considerably at another (School C), and was essentially flat at the third (School E). In implementation year 2, scores at the sole test-prep school (School B) declined, while mean scores at three of the four non-test-prep schools were essentially flat (Schools A, C, and D), and performance at School E declined. Recall that New Classrooms staff changed the content approach in the third year, eliminating the notion of test-prep rounds. In the final implementation year, Schools A and E both used the same strategy (Outcome Strategy B.2), but performance at one school increased (School E) while it decreased at the other (School A). Among the schools using the second strategy (Outcome Strategy AB.2), performance at one school increased (School B) and decreased at another (School D). As noted above, performance at School C in year three was strongly influenced by a changing student clientele.

Table 2. Test Prep Period Usage and Outcome Strategy by School and Year

	2015-16	2016-17	2017-18
School A	No Test Prep Period	Test Prep Period	Outcome Strategy B.2
School B	Test Prep Period	No Test Prep Period	Outcome Strategy AB.2
School C	No Test Prep Period	Test Prep Period	Outcome Strategy AB.2
School D	Test Prep Period	Test Prep Period	Outcome Strategy AB.2
School E	No Test Prep Period	Test Prep Period	Outcome Strategy B.2

CITS Results

This section presents the results of the comparative interrupted time series (CITS) models, our primary analyses establishing the impact of TtO on student mathematics performance. The first column in Table 3 indicates the unadjusted TtO treatment effects for each of the three treatment years, represented by the cross-level treatment year by TtO interaction terms. The direction and magnitude of the estimates here mirror exactly the V-shaped pattern of TtO school performance shown descriptively above in Figure 1, where TtO school test scores were (at least visually) above trend in year one, slightly below trend in year two, and essentially at trend in year three. Indeed, this is the picture portrayed in the first column. TtO estimates in years one and three are statistically non-significant and quite close to zero, with a slightly larger (but still statistically non-significant) negative estimate for year two. The estimates in the second column are adjusted for student academic and socio-demographic background characteristics, as well as aggregate indicators of these characteristics at the within-school cohort level. Although the adjusted TtO estimates here are slightly larger (more negative) in year two and negative in year three, all of the yearly estimates remain statistically non-significant. In sum, the results of the CITS models were statistically non-significant in each of the three implementation years.

We should also note that compared to Elizabeth Public School students in the same cohort within the same school, regardless of whether they participated in TtO, females slightly underperformed their male peers in mathematics, while language minority, special education, and free-lunch status were negatively associated with math achievement, all else equal. Asian students generally exhibited somewhat higher mathematics performance compared to white students, while black and Hispanic students had lower average test scores compared to their white peers, even on an adjusted basis. The small, positive estimate associated with reduced-price compared to full-price lunch is more difficult to explain. To capture changing demographics within EPS schools over time, these models also included aggregate indicators of these student-level measures at the cohort level. Increased proportions of LEP, special education, and black students across cohorts within schools were negatively associated with student performance, even holding the student-level associations constant.

Table 3. CITS (Difference-in-Differences) Estimates of TtO Impact on Student Mathematics Achievement in Elizabeth, New Jersey, Implementation Years One through Three

	Unadjusted	Adjusted
Students (Level 1)		
Female	--	-0.076 (0.009)***
Limited English Prof.	--	-0.739 (0.015)***
Special Education	--	-0.864 (0.016)***
Free Lunch[1]	--	-0.059 (0.014)***
Reduced-Price Lunch[1]	--	0.067 (0.020)**
Mobile	--	-0.148 (0.013)***
Black[2]	--	-0.461 (0.021)***
Hispanic[2]	--	-0.170 (0.019)***
Asian[2]	--	0.197 (0.043)***
Sixth Grade[3]	--	-0.007 (0.011)
Eighth Grade[3]	--	0.000 (0.012)
Cohorts (Level 2)		
% Female	--	0.315 (0.320)
% Limited Eng. Prof.	--	-1.443 (0.290)***
% Special Education	--	-1.387 (0.328)***
% Free Lunch	--	-0.574 (0.317)
% Red.-Price Lunch	--	-0.922 (0.526)
% Mobile	--	0.114 (0.208)
% Black	--	-1.006 (0.612)
% Hispanic	--	-0.190 (0.548)
% Asian	--	-1.709 (1.204)
Schools (Level 3)		
Ever TtO	-0.168 (0.117)	-0.165 (0.117)
Trt. Year 1		
Intercept	-0.037 (0.045)	-0.019 (0.044)
TtO	**0.067 (0.093)**	**0.062 (0.089)**
Trt. Year 2		
Intercept	0.011 (0.044)	0.026 (0.045)
TtO	**-0.082(0.091)**	**-0.113 (0.087)**
Trt. Year 3		
Intercept	-0.008 (0.044)	0.033 (0.051)
TtO	**0.030 (0.092)**	**-0.170 (0.087)**
Intercept	0.038 (0.057)	0.038 (0.057)

*$p<.05$; **$p<.01$; ***$p<.001$. Standard errors are in parentheses. Outcome is z-scored (standardized) within grades and years. Sample includes 36,158 student-level measurements nested within 209 school-by-year cohorts, within five TtO and 16 non-TtO schools.

[1] Free and reduced-price lunch compared to full-price lunch students.

[2] Racial/ethnic comparisons to white students.

[3] Grades compared to seventh grade.

Alternative Model Specifications

We also ran a series of alternative models. The first of these employed an indicator that captured the combined (averaged) three-year TtO treatment effect (see Table 4). As indicated in the first and second columns, both the adjusted and unadjusted combined three-year TtO estimates were close to zero and statistically non-significant. We constructed a second set of models in response to the dramatic change to the demographics in School C. These models removed this school from the sample. The results (not presented here) were nearly identical to those presented in Table 3, with no significant (positive or negative) TtO estimates in any treatment year.

A third set of models eliminated six control schools with pre-implementation means above the upper bound of the highest-scoring TtO school (see Table 5). As above in Table 3, the Ever TtO coefficients here represent the estimated difference in mathematics achievement between TtO and non-TtO schools *prior* to the TtO implementation. Both the unadjusted and adjusted Ever TtO estimates are non-significant, as they were with the full sample, but the estimates here are closer to zero. This suggests a somewhat better pre-treatment balance with this reduced sample. However, the TtO treatment estimates for all three years, both adjusted and unadjusted, remain statiscially non-significant.

Table 4. CITS (Difference-in-Differences) Estimates of TtO Impact on Student Mathematics Achievement in Elizabeth, New Jersey, Implementation Years One through Three Combined

	Unadjusted	Adjusted
Students (Level 1)		
Female		
Limited English Prof.	--	-0.076 (0.009)***
Special Education	--	-0.739 (0.015)***
Free Lunch[1]	--	-0.864 (0.016)***
Reduced-Price Lunch[1]	--	-0.059 (0.014)***
Mobile	--	0.067 (0.020)**
Black[2]	--	-0.148 (0.013)***
Hispanic[2]	--	-0.461 (0.021)***
Asian[2]	--	-0.170 (0.019)***
	--	0.197 (0.043)***
Sixth Grade[3]		
Eighth Grade[3]	--	-0.007 (0.011)
	--	0.000 (0.012)
Cohorts (Level 2)		
% Female	--	0.315 (0.320)
% Limited Eng. Prof.	--	-1.443 (0.290)***
% Special Education	--	-1.387 (0.328)***
% Free Lunch	--	-0.574 (0.317)
% Red.-Price Lunch	--	-0.922 (0.526)
% Mobile	--	0.114 (0.208)
% Black	--	-1.006 (0.612)
% Hispanic	--	-0.190 (0.548)
% Asian	--	-1.709 (1.204)
Schools (Level 3)		
Ever TtO	-0.168 (0.117)	-0.165 (0.117)
Trt. Years 1-3		
Intercept	-0.011 (0.029)	0.011 (0.034)
TtO X Trt. Yrs. 1-3	**0.005 (0.060)**	**-0.077 (0.059)**
Intercept	0.038 (0.057)	0.038 (0.057)

*p<.05; **p<.01; ***p<.001. Standard errors are in parentheses. Outcome is z-scored (standardized) within grades and years. Sample includes 36,158 student-level measurements nested within 209 school-by-year cohorts, within five TtO and 16 non-TtO schools.

[1] Free and reduced-price lunch compared to full-price lunch students.

[2] Racial/ethnic comparisons to white students.

[3] Grades compared to seventh grade.

Table 5. CITS (Difference-in-Differences) Estimates of TtO Impact on Student Mathematics Achievement in Elizabeth, New Jersey, Implementation Years One through Three, Reduced Sample

	Unadjusted	Adjusted
Students (Level 1)		
Female	--	-0.067 (0.012)***
Limited English Prof.	--	-0.771 (0.018)***
Special Education	--	-0.876 (0.019)***
Free Lunch[1]	--	-0.031 (0.019)***
Reduced-Price Lunch[1]	--	0.122 (0.026)***
Mobile	--	-0.130 (0.015)***
Black[2]	--	-0.450 (0.027)***
Hispanic[2]	--	-0.157 (0.025)***
Asian[2]	--	0.221 (0.053)***
Sixth Grade[3]	--	-0.014 (0.014)
Eighth Grade[3]	--	-0.001 (0.014)
Cohorts (Level 2)		
% Female	--	0.301 (0.312)
% Limited Eng. Prof.	--	-1.357 (0.305)***
% Special Education	--	-1.259 (0.325)***
% Free Lunch	--	-0.266 (0.378)
% Red.-Price Lunch	--	-0.930 (0.612)
% Mobile	--	0.120 (0.212)
% Black	--	-1.040 (0.637)
% Hispanic	--	-0.475 (0.623)
% Asian	--	-1.131 (1.292)
Schools (Level 3)		
Ever TtO	-0.019 (0.085)	-0.015 (0.086)
Trt. Year 1		
Intercept	-0.037 (0.050)	-0.053 (0.048)
TtO	**-0.019 (0.087)**	**0.102 (0.083)**
Trt. Year 2		
Intercept	0.011 (0.049)	0.068 (0.052)
TtO	**0.078 (0.085)**	**-0.136 (0.083)**
Trt. Year 3		
Intercept	-0.008 (0.049)	0.043 (0.055)
TtO	**-0.086 (0.085)**	**-0.161 (0.082)**
Intercept	0.021 (0.049)	0.022 (0.049)

**p*<.05; ***p*<.01; ****p*<.001. Standard errors are in parentheses. Outcome is z-scored (standardized) within grades and years. Sample includes 24,666 student-level measurements nested within 149 school-by-year cohorts, within five TtO and 10 non-TtO schools.

[1] Free and reduced-price lunch compared to full-price lunch students.

[2] Racial/ethnic comparisons to white students.

[3] Grades compared to seventh grade.

CONCLUSION AND DISCUSSION

This study described the effects of *Teach to One: Math* (TtO) on student mathematics outcomes during a three-year implementation in Elizabeth, New Jersey. Our primary causal analyses entailed a comparative interrupted time series (CITS) approach (also known as the difference-in-differences technique), with a sample of 36,158 student-level measurements, nested within 209 school-by-year cohorts, located within five TtO schools and a demographically and academically comparable group of 16 non-TtO schools, during the 2015-16, 2016-17, and 2017-18 schools years. We used this technique to compare changes in state-mandated, grade-level standardized test scores among TtO schools before and after program implementation to the changes in outcomes among a similar group of non-TtO schools in Elizabeth.

We found that performance patterns across the five TtO treatment schools were quite inconsistent, and the TtO estimates for each of the three implementation years were statistically non-significant. We found similar results using a measure indicating the combined (averaged) three-year impact and models that excluded a number of relatively higher-achieving control schools. Importantly, however, during the implementation period, participating TtO schools in Elizabeth asked for a variety of program configurations and alterations. Specifically, different schools requested adjustments that either emphasized or deemphasized the extent to which their students received grade-level mathematics content. This heterogeneity in both outcomes and the specific nature of the treatment itself leaves us unable to draw generalizable conclusions, positive or negative, regarding TtO's impact on student mathematics performance as measured by state-mandated grade-level assessments.

Limitations

As with any impact study, despite the use of robust methods and data, our results should be interpreted in light of limitations related to both research design and program implementation. First, these findings come from a single school district, with a particular set of teachers in a three-year implementation period. These five schools represented only about 15% of the total number of schools implementing TtO nationally during this period. TtO may well have had different results—either positive or negative—in a different context in a different point in time. Second, although we had access to over 36,000 student-level test scores and seven years of prior achievement data, the study still involved only five treatment and 16 control schools. Although adequate from an analytic perspective, our ability to identify statistically significant treatment effects that are modest in size—again, either positive or negative—was somewhat limited.

Readers should bear in mind an additional factor when interpreting these results, or indeed the results of any study of personalized learning that employs standardized test scores as outcomes. A central characteristic of TtO is that students who are missing foundational math skills are provided the time and space to experience below grade-level content. One result of this differentiation is that TtO and non-TtO students with the same initial test scores likely receive quite different mathematics content during the academic year. Because TtO students may receive content that is below grade-level, their performance on state-mandated, grade-level standardized assessments may lag behind their non-TtO peers who are being exposed to new, grade-level skills over the course of the academic year. This would be the case regardless of whether non-TtO students possess the fundamental skills that New Classrooms staff argue are needed for long-term success in mathematics. In other words, mere exposure to increased levels of grade-level content—although possibly inappropriately advanced—may provide a near-term impact on grade-level assessments. But it may not be the best approach for all students in the longer term.

A final consideration is that the TtO model continues to evolve. Its design elements and algorithms are not fixed and can presumably improve over time. As such, it is unclear the extent to which the model examined during this implementation period is representative of the TtO model currently being used in schools today. In sum, these analytic complications, which are likely common to any evaluation of any personalized learning model, obviously warrant further consideration.

REFERENCES

Barr, R., & Dreeben, R. (1983). *How schools work*. Chicago: University of Chicago Press.

Bloom, H.S. (2003). Using ‘’short’’ interrupted time-series analysis to measure the impacts of whole-school reforms: With applications to a study of accelerated schools. *Evaluation Review*, *27*(1), 3-49.

Bruner, J. S. (1978). The role of dialogue in language acquisition. In A. Sinclair, R.J. Jarvelle, & W.J.M. Levelt (Eds.), *The child’s concept of language (pp. 241-256).* New York: Spring er-Verlag.

Cole, R., Kemple, J.J., & Segeritz, M.D. (2012). *Assessing the early impact of School of One: Evidence from three school-wide pilots*. New York: Research Alliance for New York City Schools, New York University. Retrieved from http://steinhardt.nyu.edu/scmsAdmin/media/users/sg158/PDFsschool_of_one/AssessingEarlyImpactSo1_Exec.pdf

Corno, L. (2008). On teaching adaptively. *Educational Psychologist*, *43*(3), 161-173.

Gates Foundation. (2014). *Early progress: Interim research on personalized learning*. Retrieved from http://k12education.gatesfoundation.org/student-success/personalized-learning/early-progress-interim-report-on-personalized-learning-rand/d-learning-and-other-edtech-buzzwords-on-the-gartner-hype-cycle.

Lee, V.E., & Ready, D.D. (2009). The U.S. high school curriculum: Three phases of research and reform. *The Future of Children*, *19*(1), 135-156.

Martinez, J.F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgements, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment*, *14*(78), 78-102.

Newcomb, T. (2016, June 16). Teach to One: Inside the personalized learning program that Bill Gates calls the “future of math.” *The 74 million.* Retrieved from https://www.the74million.org/article/teach-to-one-inside-the-personalized-learning-program-that-bill-gates-calls-the-future-of-math

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

Pallas, A. P., Entwisle, D. R., Alexander, K. L., & Stluka, M. F. (1994). Ability-group effects: Instructional, social, or institutional? *Sociology of Education*, *67*, 27-46.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Ready, D. (2014). *Student mathematics performance in the first two years of Teach to One: Math*. New York: Center for Technology & School Change, Teachers College Columbia University. Retrieved from http://www.newclassrooms.org/wp-content/uploads/2016/09/Teach-to-One_Report_2013-14.pdf.

Rockoff, J. (2015). *Evaluation report on the School of One i3 expansion*. New York: Columbia University Business School. Retrieved from http://www.edweek.org/media/evaluation%20of%20the%20school%20of%20one%20i3%20expansion%20--%20final%20copy.pdf.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Snow, R.E. (1980). Aptitude, learner control, and adaptive instruction. *Educational Psychologist*, *15*, 151-158.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.