Growing Teachers from Within

Implementation, Impact, and Cost of an Alternative Teacher Preparation Program in Three Urban School Districts—Technical Appendix

JULIA H. KAUFMAN, BENJAMIN K. MASTER, ALICE HUGUET,
PAUL YOUNGMIN YOO, SUSANNAH FAXON-MILLS,
DAVID SCHULKER, GEOFFREY E. GRIMM

Sponsored by TNTP

RAND
CORPORATION

For more information on this publication, visit www.rand.org/t/RRA256-1

Published by the RAND Corporation, Santa Monica, Calif.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Preface

This technical appendix details research methods and tables for the impact analysis of an investigation of implementation of the Teacher Effectiveness and Certification (TEACh) initiative in three urban school districts.

This study was undertaken by RAND Education and Labor, a division of the RAND Corporation that conducts research on early childhood through postsecondary education programs, workforce development, and programs and policies affecting workers, entrepreneurship, and financial literacy and decisionmaking. TNTP (formerly The New Teacher Project) sponsored the report, with funding through a U.S. Department of Education Supporting Effective Educator Development (SEED) grant program, grant number U367D150008. TNTP is committed to ending educational inequities by promoting the recruitment, training, and retention of high-quality teachers and school leaders. SEED has funded TNTP to implement the TEACh initiative in three urban school districts.

The authors would like to acknowledge the following people for their contributions to this work. First, our thanks to Serafina Lanna, Katie Whipkey, and Naomi Hale, who made considerable contributions to the data collection and analysis for this work at RAND. Second, thanks to TNTP for support and input over the course of this research. We would especially like to thank Cassandra Coddington for her regular guidance and feedback, as well as Ellie Leahy, Bailey Cato Czupryk, Genine Blue, and Camilo Gonzalez. We would also like to thank the TNTP staff in each of the three participating districts who were so collaborative and helpful in providing us with data and advice related to our analysis, including Caitlin Deschenes-Desmond, Jessica Dirks, Heather Barondess, and Erin Martin.

More information about RAND can be found at www.rand.org. Questions about this report should be directed to jkaufman@rand.org, and questions about RAND Education and Labor should be directed to educationandlabor@rand.org.

Appendix A. Methods

We relied on a mixed-methods approach to gather and analyze data related to the TEACh program.

Implementation Study Methods

Approach

Data for the implementation portion of this report were collected through six research activities: (1) interviews with TNTP and district staff; (2) focus groups with teaching candidates during their pre-service training (PST) experience; (3) observations of PST training sessions; (4) phone interviews with teaching candidates in the spring of their first year in the classroom; (5) phone interviews with principals who hired program teaching candidates, also in the spring of the candidates' first year in the classroom; and (6) document collection from program staff. These data collection activities were designed with a set of implementation measures in mind, identified by TNTP in its original proposal and later modified through a collaboration between TNTP and researchers at the RAND Corporation (see Table A.1).

Table A.1. Implementation Measures

1	**Hiring Metrics**
1a	A diverse set of recruitment activities is in place.
1b	Hiring and selection criteria and processes are in place and are clear and transparent.
1c	Districts allocate sufficient personnel and resources to recruitment activities and processes.
1d	School leaders have clear understanding of and buy-in to program.
1e	Program recruits and trains teachers to serve at placements in district's high-need certification areas.
2	**Teacher Effectiveness**
2a	District staff responsible for PST and coaching are in place and prepared for their roles.
2b	PST is focused on skills considered essential to the district and program.
2c	PST provides considerable opportunities to practice teaching in a classroom setting.
2d	In-service coaching provides regular, specific feedback on changes to instructional practice.
2e	District and program evaluation systems provide feedback to teachers and differentiate according to performance.
3	**Program Sustainability**
3a	There are plans, processes, and resources in place to transition the program from TNTP to the district.

Data Sources

Two RAND team members visited each program for several days in the summer and fall for two consecutive years. During both the summer and fall visits, we conducted interviews with TNTP and school district staff (including coaches and training instructors). During the summer visits only, we also led focus groups with candidates currently being trained in the program and observed several hours of PST sessions, taking notes as relevant to the implementation measures. We collected documents from program staff, such as *stepbacks* (reflections on program progress developed by TNTP for each internal team), training content, and enrollment data. Each spring we conducted phone interviews with program candidates who were in their first year of teaching, as well as phone interviews with principals who had hired TEACh program candidates that school year.

We selected district and TNTP staff for interviews because of their association with the program. All candidates enrolled in PST were invited to attend the focus groups, and between four and ten participants attended each of the PST focus groups we conducted. We randomly selected candidates for spring phone interviews and invited them to interview via email. We invited the principals of candidates to participate, also via email. In the case of the PST focus groups and candidate and principal phone interviews in Districts A and B, RAND compensated participants for their time with a $20 Target gift card or Amazon gift code following the interview. In District C, at the district's request, we instead provided the district $20 per interviewee to go toward a district professional development fund.

Table A.2. Implementation Study Interview Counts, by Year

	District A	District B	District C
2016–2017 school year			
District staff interviews	6	6	0
TNTP staff interviews	5	3	0
Coach interviews	3	2	0
PST candidate focus groups	3	2	0
Candidate interviews	8	8	0
Principal interviews	3	3	0
2017–2018 school year			
District staff interviews	6	5	4
TNTP staff interviews	9	5	5
Coach interviews	5	4	7
PST candidate focus groups	2	2	2
Candidate interviews	9	8	8
Principal interviews	3	2	4
2018–2019 school year			
District staff interviews	0	0	12
TNTP staff interviews	0	0	5
Coach interviews	0	0	3
PST candidate focus groups	0	0	2
Candidate interviews	0	0	8
Principal interviews	0	0	4

We wrote interview protocols in alignment with the project's implementation measures (see Table A.1). Interviews with district staff and PST focus groups lasted approximately one hour, while interviews with candidates and principals were each 30 minutes. In the 2016–2017 and 2017–2018 school years, RAND team members took transcript-style notes during interviews, then listened to interview audio to check the notes for accuracy. In the 2018–2019 school year, we instead sent interview audio to professional transcribers. In all years, we uploaded notes and transcripts into the qualitative software Dedoose for coding.

Analysis

We developed the coding scheme to align with the study's implementation measures, outlined in Table A.1. We applied deductive codes that were descriptive (e.g., we identified when an interviewee was discussing *recruitment*), as well as analytic (e.g., we identified when the discussion was about *strength* of the program). We also remained open to emergent codes that arose as we reviewed our transcripts and notes. For example, if coders saw that a topic was discussed by multiple interviewees, but we did not have an existing code for it, we would create a code for that topic. In each round of coding, a group of three or four coders practiced coding on the same transcripts, followed by team debriefs to identify, discuss, and correct any areas of misalignment. After establishing consensus on codes for several transcripts coded by multiple individuals, coders coded transcripts individually.

During analysis, we focused on areas of successes or concerns that were shared across programs, as well as identifying when a unique issue in a single program was mentioned by multiple interviewees. We organized findings by specific program components in our report: (1) recruitment and selection, (2) PST, (3) hiring into school-year positions, (4) on-the-job support, (5) evaluation, and (6) program sustainability. We triangulated emergent themes by reviewing multiple interviewees' input, as well as by looking across documents and observation notes.

Impact Analysis Methods

The TEACh initiative had multiple goals, including recruiting new teachers from an alternative pipeline to fill hard-to-staff teaching positions, ensuring that those teachers were highly effective, and encouraging highly effective TEACh teachers to remain in the district. We performed descriptive analyses to summarize the program's contributions to the districts' teacher recruitment pipelines. These analyses are fully described in the main study report.

Additionally, we conducted statistical comparisons of effectiveness (as measured by student achievement and teacher evaluations) and retention between TEACh program participants and similarly situated teachers who did not participate in the program. Some of our outcomes, including student achievement in English language arts (ELA) and mathematics and teacher retention, were reasonably equivalent across district contexts. For these outcomes, our primary focus was on tests of the pooled TEACh effect across multiple districts. When pooling results across districts in this way, we used a fixed-effects meta-analysis method in which each district's estimate was weighted proportional to its precision (Cooper, Hedges, and Valentine, 2009).[1] However, teacher evaluation outcomes were very different across districts. We thus presented evaluation outcomes estimates separately for each district and did not pool them together.

Next, we detail the methodology for our comparative analyses.

TEACh Impacts on Student Achievement

To address our research questions related to the effect of TEACh program teachers on student achievement, we compared achievement outcomes for the students of TEACh teachers

[1] We opted to use a fixed-effects framework because our primary goal was to measure the average effect of the grant-funded intervention in these particular districts, and because we did not have an a-priori expectation that effects would differ across districts. Further, with a sample of just three districts, each unique, we do not have a particularly useful sample for estimating a random-effects estimate that would attempt to speak to the generalizability of our findings across a national or other population of interest.

relative to outcomes for the students of other teachers hired in the same year in each district. We refer to these teachers as *first-year* teachers if they were in their first year of working as full-time teachers in their district, and as *second-year* teachers when comparing outcomes for TEACh and non-TEACh teachers in the second year following initial hire in their district. We could not, however, observe whether some first-year teachers (either TEACh or non-TEACh) might have had any prior full-time teaching experience in a different district.

We focused specifically on students taught in classrooms by either TEACh or comparison teachers in a course directly related to the test-score outcome (i.e., either ELA or mathematics). Any student taught by both a TEACh teacher and a non-TEACh teacher from our comparison group in the same subject and year was excluded from our analysis.[2]

When comparing achievement outcomes, we employed statistical techniques to help ensure that students of TEACh and non-TEACh teachers were comparable to each other so that we could better assess whether TEACh program teachers were, in fact, more effective. Specifically, we used a method known as *doubly robust estimation* (Bang and Robins, 2005; Kang and Schafer, 2007) to compare the students of TEACh teachers with a statistically adjusted group of comparison students taught by comparison teachers hired in the same year, where the adjustment was designed to remove differences in the other factors in the data that might affect achievement. The name *doubly robust* arises from the fact that the procedure has two steps, each of which affords an opportunity to correct for the differences in other achievement-related factors that muddy the group comparison, conditional on some assumptions. One component of the method involves weighting each comparison student according to how similar he or she was to the students of TEACh teachers; another component attempts to account for these characteristics as control variables in a linear regression model.

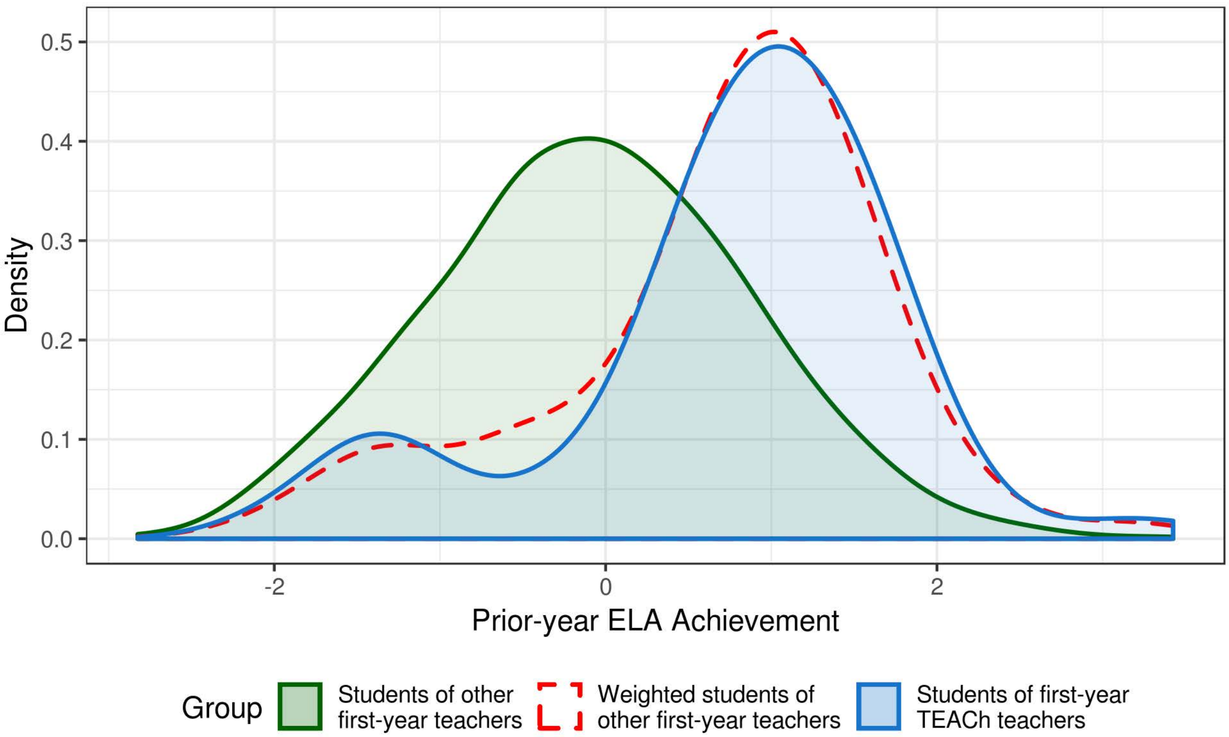
To provide a weight for each comparison student, we relied upon a similarity metric known as the *propensity score* (Rosenbaum and Rubin, 1983), which is the conditional probability of being a student of a TEACh teacher given a student's observed characteristics. Comparison students who are more similar to the TEACh students will tend to have higher propensity scores and, thus, receive more weight. When all comparison students are weighted appropriately, they should match the students of TEACh teachers closely on the characteristics that factor into the estimated propensity scores. To estimate the propensity score for each comparison student, we used a machine learning technique known as a generalized boosted model (GBM; Ridgeway, Madigan, and Richardson, 1999). GBMs are flexible and nonparametric, and they iteratively build a model of the relationship between the observable characteristics and the type of new teacher that each student was assigned to. The model accounts for the possibility of complex interactions and nonlinearities. This flexibility makes GBM a powerful tool for finding

[2] A relatively small number of students was linked to multiple TEACh or comparison teachers, and the fact that we considered multiple cohorts in some analyses created the possibility that students could be considered treated in one analysis and untreated in a separate analysis. To eliminate the possibility of attributing the effects of the program to the control group, we removed students linked to a TEACh teacher from all comparison groups in each school year.

the set of propensity scores that minimizes the group differences in control variables (Lee, Lessler, and Stuart, 2010).[3]

Figure A.1 provides an illustrative example of the weighting aspect using the comparison of students who had either TEACh or other first-year teachers. Figure A.1 compares the prior-year ELA achievement of the two groups of students using a density plot, which is a smoothed estimate of the distribution of scores in each group across the spectrum of score values. The density plot shows that, in this example, students of TEACh teachers tended to have higher ELA achievement in the year prior to our analysis year than comparison students. The red dashed line, which represents the comparison students after weights are applied, demonstrates that the weighted control students more closely resembled the TEACh students subsequent to weighting.

Figure A.1. Illustrative Example of Group Differences in Prior-Year ELA Achievement



Differences such as those in Figure A.1 could theoretically exist across many different dimensions. Table A.3 presents all the available information that we include in the adjustment procedure for analysis of student achievement outcomes. The procedure adjusted the comparison

[3] We implemented the doubly robust procedure using a modified version of the fastDR program in R (Ridgeway, 2014). We used a minimum of 20,000 GBM iterations in each propensity score model, and we measured the balance at each iteration according to the Kolmogorov-Smirnov (KS) statistic, which is a nonparametric measure of the similarity of two distributions. After running the full model, fastDR searches for the iteration that produces the optimal balance (i.e., similarity between students of TEACh teachers and comparison students).

students to mirror the students of TEACh teachers on a variety of demographic and achievement metrics to help ensure that remaining test score differences between the groups were more plausibly attributable to the differential impact of TEACh teachers.

Ideally, the weighting procedure would remove all observed differences between the comparison students and the students of TEACh teachers, but the doubly robust method takes the additional step of including the weights in a regression model that also includes the characteristics as regression controls with a standard additive linear specification. In Table A.3, we provide a detailed summary of the variables included in our models for estimating student achievement impacts. The inclusion of regression controls provided an additional layer of safety in case the weighting did not fully remove the differences, but the risk of misspecification in the linear model was lower because the two groups of students were made similar via the weighting. We implemented the procedure with a weighted ordinary least squares regression.

Table A.3. Variables Included in the Doubly Robust Student Achievement Comparisons

Category	Variable	Included in Weighting	Included in Regression
Student demographics	Grade level	Yes	Yes
	Race/ethnicity	Yes	Yes
	Gender	Yes	Yes
	Free/reduced price lunch[b]	Yes	Yes
	English Language Learner (ELL)	Yes	Yes
	Special needs	Yes	Yes
	At risk of dropout (district-defined)[b]	Yes	Yes
	Mother's level of education[a]	Yes	Yes
Student prior achievement and behavior	Prior-year ELA test score	Yes	Yes
	Prior-year math test score	Yes	Yes
	Number of disciplinary actions	Yes	Yes
	Attendance	Yes	Yes
Other TEACh contextual factors	Specific role distinctions linked to special needs teachers[a]	Yes	Yes
	Linked to a designated ELL specialist teacher[a]	Yes	Yes
	Percentage of other students of associated teacher who are ELLs[a]	Yes	Yes
	Whether associated teacher taught exclusively special needs students, non–special needs students, or a mix of both[a]	Yes	Yes
School characteristics	Average test scores of students in previous year in ELA and mathematics[c]	No	Yes
	Percentage of students eligible for free/reduced price lunch[b]	No	Yes
	Percentage of students who are black or Hispanic[a]	No	Yes

NOTES: [a] indicates variables that were not included in the District B comparisons.

[b] indicates variables that were not included in District C comparisons. Models that included multiple cohorts also included an indicator for school year.

[c] We calculated this variable by first creating school-wide average values for the standardized mathematics and ELA test scores (normalized as z-scores), and then averaging the mathematics and ELA school-wide values for each school.

We did not include characteristics of the teachers themselves, such as certifications or demographics, in the doubly robust procedure. Teacher characteristics could certainly contribute to differences in student test scores, but we chose to exclude them so that any peculiar characteristics of the TEACh teachers would be captured in the estimates of the program impacts. If TEACh teachers were more likely to earn key certifications, for instance, and those certifications made them more effective, then controlling for this difference would not allow it to contribute to the program effect. Instead, our aim was to compare TEACh teachers to other teachers that the district hired, without adjusting for any contributions related to TEACh teachers' training or other teacher characteristics. As Table A.3 shows, we did, however, include some variables that related to teachers' job roles, as these could help us to better distinguish important nuances in the types of students taught by each teacher (e.g., between special education students who are moderately versus severely challenged).

A relatively small number of students were missing values for some of the covariates, so we used a multiple imputation procedure described in Honaker and King (2010). Multiple imputation fills in missing values using a predictive model that incorporates all of the known information about each particular student. Importantly, this procedure avoids excluding any students from the analysis, while also capturing the additional uncertainty introduced by the incomplete data by generating several imputations for each missing value and then aggregating the results.

A key aspect of any statistical method for measuring group differences is the estimation of the amount of uncertainty surrounding the effects. Standard approaches can calculate the amount of uncertainty using established formulas, but there are no such formulas for machine learning algorithms such as the GBM approach that we used to estimate the propensity scores. Thus, we turned to resampling methods to account for all of the variability in the entire procedure, including the GBM process of estimating the propensity scores (McCaffrey, Ridgeway, and Morral, 2004). Resampling methods capture the uncertainty in the estimation process by repeating the estimation process in its totality (which, in our case, includes multiple imputation,[4] propensity score estimation, and regression estimation) on many random combinations of actual students in the data set. Specifically, we used a nonparametric bootstrap that was designed to allow for potential residual correlation within schools (Cameron and Trivedi, 2005, p. 845).[5]

To summarize, the full estimation procedure involved the following steps. First, we generated 100 bootstrap replications for each district by sampling schools with replacements until the total number of schools in the bootstrap sample equaled the district total. Then, we

[4] Combining a bootstrap procedure with multiple imputation raises the question of whether one should impute the missing values before or after generating the bootstrap data sets. We chose to repeat the multiple imputation procedure on each bootstrap data set based on the findings in Schomaker and Heumann (2018).

[5] Cluster adjustments could be unreliable if there are too few clusters (Cameron and Miller, 2015). This was not a problem for the overall estimates, but for subgroup models, the number of clusters in a subgroup was occasionally borderline. All of the statistically significant results we report were in samples that included at least 30 clusters. Also, we calculated significance using a t-distribution, with the degrees of freedom based on the number of clusters (i.e., schools) rather than the number of students.

created five imputed data sets for each of the 100 replications. We estimated doubly robust treatment effects for each imputed data set and averaged these effects to produce an estimate for each bootstrap replication. Then, we calculated the standard error of the treatment effect as the standard deviation across the 100 bootstrap estimates.

Table A.3 points to one additional limitation in the analysis regarding how we accounted for school characteristics. Ideally, we would weight on school-level factors in the same way as the student characteristics, thus creating a weighted comparison group made up of students attending similar types of schools. For our analyses, the inclusion of school characteristics in the weighting generally degraded the balance achieved on the student characteristics. The reason for this is that the school-level controls incentivized the model to add more weight to comparison students in schools with similar test scores or demographics to the schools that hired TEACh teachers, but some of the students in those schools were less comparable to TEACh students in other ways, putting the balance in the school characteristics at odds with the student characteristics. Our preferred option, in light of this limitation, was to use the weighting to create a comparison group of non-TEACh students across the district, but to rely on the regression models to control for the relationship between school characteristics and test scores.[6] Fortunately, as shown in Appendix B, Tables B.12, B.13, B.15, and B.16, school-level characteristics were not very different overall between TEACh and comparison teachers. Still, the reader should bear this limitation in mind when thinking conceptually about the comparison students.

This appendix provides a full description of the similarities between the students of TEACh teachers and comparison students, before and after weighting, in each district sample. Tables in Appendix B illustrate the effectiveness of the weighting in the student achievement analyses by providing the average levels of each characteristic, the differences in group means expressed as effect sizes, as well as the Kolmogorov-Smirnov (KS) statistic, which is a metric that captures the similarity between the groups across the entire range of values of the characteristic. The tables show that the weighting procedure successfully removed major differences in the student-level characteristics (such as those in Figure A.1), which is reflected in the similar average levels between the “Adjusted” and “TEACh” columns, as well as the substantial reductions in the KS statistics before and after weighting.

Finally, our analyses still relied on the assumption that there were no systematic differences between TEACh and comparison students after accounting for the other observable characteristics available in the data. This design leaves open the possibility that other systematic differences that may exist between the groups and could not be directly observed may introduce

[6] To test the sensitivity of the results to this decision, we also examined a specification that broke each school metric into thirds and incorporated this discrete metric into the propensity score model. This specification sacrifices some information in the school metric, but enabled us to incorporate the school metric into both stages of the analysis. The results were broadly consistent across these two specifications. In addition, descriptive plots showed that the relationship between school characteristics and student test scores was approximately linear in our data, further suggesting that the linear school-level controls are adequate to account for these factors.

bias into the group comparisons. As a purely hypothetical example, our results could have been biased if TEACh teachers were more likely to get hired in schools with less experienced principals. This could impact their students' performance, but we would not be able to control for this difference in their assignments because we do not have data on principals' experience levels. This limitation aside, the comparison of observationally equivalent students in each group provides the best information available on the relative impacts of TEACh teachers on student learning.

TEACh Impacts on Teacher-Level Outcomes

To address research questions focused on teacher evaluations and retention, we conducted a similar doubly robust comparative analysis between TEACh teachers and comparison teachers hired in the same school year. We considered teacher-level professional outcomes including teachers' summative job evaluation ratings provided by their district in each year and teacher retention in the district. A teacher was defined as retained if s/he was present in the district data file in one school year and in the subsequent school year.

As before, direct comparisons between TEACh teachers and others could be misleading if there were systematic differences between the two groups in terms of where or whom they teach. To control for these potential differences, we applied the same estimation procedure used in our student-level analyses to adjust the group of non-TEACh teachers to mirror the observable characteristics of the TEACh teachers before comparing their outcomes. In District B and District C, where we had reliable student-teacher links for almost all teachers in our sample, we included the average characteristics of the students assigned to each teacher in their classrooms in both the weighting and regression steps of our doubly robust estimation. However, in District A, reliable records of student-teacher links were not available for all teachers and are thus not included.

In Districts C and A, we also accounted for specific teacher job roles, including whether the teacher taught bilingual education or special education classes, when these data were available. However, in District B, where these role distinctions were not available, we relied solely on the student-level special education and ELL indicators to account for classrooms of different types. As in our student achievement analysis, we excluded other teacher characteristics, such as demographics and certifications, from our models because they could potentially contribute to the program effect that we were seeking to estimate. Table A.4 provides a complete rendering of the control variables included in the doubly robust comparisons of teacher-level outcomes, with indications of variables included in certain districts only and variables included in all districts' models. Appendix B, Tables B.11 and B.14 illustrate differences between TEACh and comparison teacher samples before and after weighting in the pooled teacher retention analyses.

Table A.4. Variables Included in the Doubly Robust Teacher Outcome Comparisons

Category	Variable	Included in Weighting	Included in Regression
Student Characteristics			
(Most Common Value)	Grade level taught	District B	District B
(Average/ Percentage)	Prior-year ELA/math scores[a]	District B, District C	District B, District C
	Race/ethnicity (black/Hispanic)	District B, District C	District B, District C
	Free/reduced lunch	District B	District B
	ELL	District B, District C	District B, District C
	Special needs status	District B, District C	District B, District C
	At risk of dropout	District B	District B
	Prior-year attendance	District B, District C	District B, District C
	Number of disciplinary actions	District B, District C	District B, District C
Teacher Characteristics	Bilingual education job role	District A, District C	District A, District C
	Special education job role	District A, District C	District A, District C
School Characteristics	Average test scores of students in previous year	District A	All districts
	Percentage of students on free/reduced lunch	District A	District B, District A
	Percentage of students who are black or Hispanic	No districts	District C
	School level (elementary, middle, or high school)	District A, District C	District A, District C

NOTE: Columns indicate which variables are included in analyses as a function of district. Analyses in District C accounted for multiple different categories of special education job roles. Models that included multiple cohorts also included an indicator for school year.

[a] Prior-year test scores were included only in the subgroup models, in which all students were in grade levels that had tested in the previous year.

We implemented the doubly robust procedure with a weighted linear regression for evaluation outcomes and a weighted logistic regression in the case of the retention outcome. For the logistic regression, we took the additional step of converting the treatment effect to the more intuitive probability scale. We used the same multiple imputation procedure described in the student achievement analyses section to impute any missing covariates,[7] and we followed the same bootstrap procedure to estimate the uncertainty surrounding each treatment effect for statistical significance determinations. As in the student-level analyses, the bootstrap procedure also accounted for potential clustering at the school level.

District Teacher Evaluation Measures

Teacher evaluation outcomes varied substantially across districts, with implications for our analytic models and for our interpretation of any results. First, as described in Table 3.1 of the main report, the district evaluation systems relied on different components. Two of the district evaluation scores that we analyzed were based entirely on principal observations and discretionary ratings. In District B, principal discretionary ratings were the primary, but not the

[7] As in the student-level analysis, we did not impute missing outcome measures; e.g., teachers who did not receive evaluations have missing values for the corresponding outcomes measures.

only, factor in the teacher evaluation ratings that we analyzed, contributing anywhere from 50 to 80 percent of rating scores. While principals may have drawn upon a variety of locally collected data in generating holistic ratings, we do not have access to those data. In District B, evaluations also incorporated formal district-wide measures drawn from surveys of students' experiences in teachers' classrooms, in grade levels 3 to 12. Also in District B, evaluations incorporated teacher's value added to student achievement in grade levels to 8 or measures of students' accomplishment of defined learning goals in other grades.

Teacher evaluations in District C differed from those in the other two districts in that they were not expected to be conducted annually for all teachers. Instead, most teachers were supposed to receive at least one "full" evaluation sometime within their first two years on the job (in practice, 82 percent did). A small minority of teachers in this district also received different types of evaluations on a targeted basis and at the discretion of their principal, but we did not include these additional evaluation types in our analyses. In theory, because evaluations in this district were not automatic, there could have been differences in the rates at which TEACh and comparison teachers were evaluated, or in which types of teachers from each group were evaluated. However, in practice, we observed that the same or nearly the same proportion of TEACh and non-TEACh teachers received a full evaluation in year one (44 percent for both TEACh and non-TEACh teachers) and in year two (82 percent versus 80 percent), which suggested to us that differential sorting of TEACh teachers to receive evaluations was unlikely to be occurring.

Across all of our teacher evaluation measures in the three districts, we opted not to standardize teacher's ratings. In part, this was because the ratings were on a holistic scale and might be more interpretable in that original scale. In addition, in District A in particular, there was so little variation in teacher evaluation ratings (i.e., almost everyone got the same rating) that standardizing ratings might have made any effect estimates more difficult to interpret.

Accounting for Multiple Performance Outcomes and Multiple Comparisons

As part of our impact evaluation, we considered a variety of different performance outcomes for TEACh teachers. For each outcome, we identified a single analysis that offered the best test of TEACh impacts on that outcome. In the findings of our report, we focus on year two impacts first, given that year two outcomes reflected TEACh performance after teachers had fully completed the program and were, thus, the most important measure of TEACh's longer-term impacts. However, year one outcomes addressed a relevant secondary question of TEACh performance while the program was ongoing and before any selective screens were applied to encourage less effective teachers to leave the district, and we thus share separate analyses of year one outcomes, when appropriate, throughout the report.

Our outcome domains and corresponding primary analytic samples were as follows:

- When analyzing TEACh effects on student achievement in mathematics or ELA in year one or year two, our test of impact was the pooled effect across districts in each case.
- When analyzing TEACh effects on teacher evaluation ratings after year two or year one, we considered each district's evaluation measure as a distinct outcome, and our test of impact in each case was the individual district-specific result in each case.
- When analyzing TEACh effects on teacher retention after year two or year one, our primary test of impact was the pooled effect across districts in each case.

We also conducted additional analyses to help aid in hypothesis generation around possible variation or trends in impacts across individual districts, over time, across cohorts, and to explore follow up analysis related to specific subgroups of interest. Conducting multiple subgroup tests within outcome domains increased the likelihood of detecting false positive results by chance. For any of these exploratory results that were significant at traditional thresholds ($p < 0.05$), we applied a Benjamini Hochberg comparison adjustment with a false discovery threshold of 10 percent (Benajimini and Hochberg, 1995). None of these findings was robust to this adjustment, and thus none is described as statistically significant in our main report.

Limitations

Our analyses of the impacts of TEACh had several important limitations, to some of which we have already alluded. First, neither our descriptive analysis of the numbers and types of teachers hired through the program nor our comparative analyses of the performance of TEACh teachers directly evaluates the impacts the program may have had on the overall teacher recruitment pipeline in each district. In theory, recruiting additional teachers who would not otherwise have gone into teaching might be expected to increase the total pool of teacher applicants, allowing districts to be more selective when filling roles, particularly in the hard-to-staff areas that the TEACh program focused on. However, we lacked sufficient historical data on job applicants, district hiring decisions, and fill-rates for job vacancies to adequately assess any potential impacts on the overall supply of teachers in each district. This represents an entire category of potential impact that is only indirectly addressed by our available data and analyses.

Second, our doubly robust statistical methodology, while rigorous, does not eliminate the possibility of bias when comparing TEACh teacher performance outcomes (of all types) to those of comparison teachers. First, in some cases it was difficult to identify ideal matches for teachers or students, in which case we relied more heavily on covariate controls to adjust for differences between dissimilar comparison groups (for example, as shown in Appendix B, Table B.11, TEACh teachers were more often elementary school teachers than weighted comparison teachers, and we account for this lingering difference using covariate controls for grade level). This was specifically a challenge in District C, in which TEACh teachers were frequently special education teachers, for whom ideal matches were difficult to identify. Second, TEACh teachers, their school and job assignments, and the students that they taught may have been different in unobserved ways that were not fully reflected in the available administrative data, and this could have led their professional performance to appear better or worse than comparison teachers in ways that were unrelated to their own performance.

A third category of limitations to keep in mind is related to the relatively small sample sizes that we analyzed in this research. In many analyses, our statistical power to detect effects was limited, and our confidence intervals were quite wide. As a result, it is entirely possible that we failed to detect true differences in TEACh performance that may have been small or moderate in size in some analyses. This was particularly the case for subgroup and cohort-specific analyses. In addition, not only were our samples small, but they varied in terms of their representativeness. In particular, our analysis of student achievement outcomes was limited to teachers of students in grades 4 through 8 who taught in the tested subjects of ELA and mathematics. This group may or may not have been typical of all TEACh teachers with respect to impacts on student learning. In addition, we were unable to estimate effects on student achievement in one of the three districts, although the TEACh program in District A was substantially smaller than in the other two districts. As a consequence of these challenges, our achievement impact results represent a potentially incomplete view of effects on student learning across all TEACh teachers.

Our student achievement analyses also relied on imperfect data detailing the connections between teachers, students, and our achievement measures. Particularly in middle school grades, teachers often taught multiple courses, and these courses likely varied in terms of their relevance to mathematics or ELA learning as measured by state tests. We intentionally used a broad lens to connect teachers to students in all courses that, based on course names, might have had a connection to each achievement outcome. This could, however, have attenuated any impacts if the curriculum in some courses were more directly aligned with state test measures than others. Identifying the most appropriate teacher-to-student links is a common challenge in analyses relating student achievement gains to teachers, but it is important to keep in mind when interpreting our results.

Analyses conducted in District A generally involved very small samples of teachers. As a consequence, as detailed in Tables A.3 and A.4, in our analyses in District A we relied on models with very few covariate controls. As a consequence, we are more concerned about our ability to create fair comparisons between TEACh teachers and comparison teachers in District A.

By design, our analyses focused on average impacts of the intervention across the three districts in which TEACh was supported by the grant funding. However, it is not clear to what extent our findings would generalize across a larger population of districts, either nationwide or even for a population of large urban districts like those in this study. While we can, with some confidence, estimate the impacts of TEACh in these particular districts, these results may or may not generalize across a larger population of districts nationwide.

Finally, the span of this study limited our ability to analyze variation in TEACh teachers' performance outcomes over time and as teachers became more experienced. We were only able to view teachers' performance in year three for one cohort of teachers in one district, and as a result we have incomplete information about how TEACh teachers may have performed over the long term. We were similarly limited in our ability to evaluate any changes in the effects of the TEACh program itself over time, given that, for many cohorts, we could not evaluate performance outcomes beyond teachers' first year on the job (i.e., prior to their completion of the program). Overall, our findings indicate whether TEACh teachers were initially more effective than other novices but do not address whether TEACh teachers were more effective than comparison teachers over the full length of their teaching careers.

Cost Study Methods

The cost study focused on (1) the annual costs of the TEACh pipeline overall and per teacher hired; (2) the costs of district and TNTP staff time to develop and run the TEACh program; and (3) the costs of developing and running various components of the TEACh program, including recruitment, pre-service preparation, hiring, and support of teachers. The cost study also compared the costs of recruiting and hiring teachers for TEACh and for traditional hiring pipelines.

Cost Study Approach

We designed the cost study to provide clear information about the resources and expenditures associated with the TEACh program. We used an *activity-based approach*—sometimes called *cost ingredients method*—in our data-gathering and analysis. Similar methods have been used to characterize school district expenditures on teacher professional development (e.g., Chambers, Lam, and Mahitivanichcha, 2008), as well as costs of pipelines that prepare, hire, and support principals (Kaufman et al., 2017).

This activity-based approach started with a list of major activities or *cost ingredients* that could be part of the TEACh program in any of the three participating districts, which was developed in collaboration with TNTP staff in all TEACh program districts. This activity list guided our data-collection process and can also serve as a resource for organizations and districts to consider what major activities are part of programs like TEACh. Table A.5 provides an overview of those activities, organized by major TEACh program component. After developing this activity list, we used it to query TNTP and district staff about the percentage of their time spent on each TEACh activity in the first three years of the TEACh program, corresponding to the period during which the SEED grant provided support to each district. One important note is that the activity list represents a slightly different set of activities from those described in our implementation analyses. Specifically, the implementation analyses examined pre-service programming separately from first-year supports, whereas the cost analysis grouped pre-service programming and first-year supports together, given the difficulty in separating out costs because the same staff often provided both pre-service and school year supports. In addition, the cost study has separated out systems and capacity as an additional component.

Table A.5. TEACh Program Activities

Program Component	Activities
1. Recruitment and selection for TEACh PST	Develop and/or revise strategic plans and systems for recruiting and selecting pre-service teacher candidates
	Recruit candidates for PST
	Screen and select candidates for PST
2. Selection, hiring, and placement in teaching positions	Develop and/or revise strategic plans for selecting, hiring, and placing candidates into school-year teaching positions
	Screen, hire, and place candidates into teaching positions
3. Pre-service and new-teacher training and support	Develop and/or revise strategic plans for preparation and support
	Deliver PST prior to teacher hiring and placement
	Provide pre-service coaching prior to hiring
	Provide on-the-job coaching, support, and evaluation for first-year TEACh teachers
	Support and/or administer teacher certification examinations or assessments
4. Systems and capacity for supporting TEACh program	Train those who support TEACh teachers
	Manage and analyze TEACh program data
	Oversee and sustain TEACh program
	Apply to be a state-approved teacher certification program

Readers should keep in mind that every district was not engaged in exactly the same activities each year. Table A.6 notes the activities taking place in each district and in each year. Some points about the variation in activities across districts and years:

- **Differences in who pursued state approval.** As shown in the table, the activity of applying to be a state-approved teacher certification program is applicable only in Districts A and C; District B's TEACh program existed and was already approved as a state program before the start of the SEED grant, which supported TNTP to work with District B to revise the existing program.
- **Differences in when cohorts were trained, hired, and coached.** In addition, in years two and three, District B began offering PST during the school year. Thus, the second cohort of District B TEACh candidates was trained in the fall of year two, with candidates hired in schools starting in in the winter of year two. While District B had one more summer PST at the end of year two, it shifted in year three to offering PST only during the year.
- **Differences from year to year.** Because year one was a planning year in each district, PST was offered only toward the end of that year in the summer in all districts, and the yearlong support for the first teaching cohort in all districts did not occur until year two, when the first cohort had been trained and was hired by schools across the districts.

Thus, the year-to-year changes in costs do not necessarily imply that costs for the same activities increased or decreased. Instead, changes in costs imply different activities being undertaken from year to year. In this report, we thus discuss average annual costs across years more than year-to-year changes in costs to give readers a better overall picture of costs and allow for more comparability between districts. When year-to-year differences are shared, readers should exercise caution in interpreting the meaning of these shifts in costs.

Table A.6. TEACh Activities in Each District and Year

	District A			**District B**			**District C (Later start)**		
	Year 1 Cohort	**Year 2 Cohort**	**Year 3 Cohort**	**Year 1: Cohorts 1a and 1b**	**Year 2: Cohorts 2a, 2b, and 2c**	**Year 3 Cohort**	**Year 1 Cohort**	**Year 2 Cohort**	**Year 3 Cohort**
2015–2016 school year	▨			▨					
2016 Summer	Trained			1a: trained					
2016–2017 school year	Hired and coached			1a: hired and coached 1b: trained in fall and hired/coached in spring			▨		
2017 summer		Trained			2a: trained		Trained		
2017–2018 school year		Hired and coached			2a: hired and coached 2b: trained in fall, hired and coached in spring 2c: trained		Hired and coached		
2018 summer			Trained					Trained	
2018–2019 school year			Hired and coached		2c: hired and coached	Trained in fall, hired and coached in spring		Hired and coached	
2019 summer									Trained
2019–2020 school year									Hired and coached

NOTE: Gray shading notes initial planning period for each district.

Cost Study Data

Five major data sources were used to gather cost information for this cost study:

- **SEED grant expenditure reports.** TNTP central office staff provided us with fiscal year expenditure reports of SEED spending by district, including average salaries for full-time categories of personnel and part-time staff hours and hourly rates. These reports also included fringe benefit rates and nonpersonnel costs for travel, supplies, and other contractual expenditures.
- **Staff percentage effort.** TNTP central office staff also provided us with information about the percentage of total time that TNTP staff members in each site reported spending on SEED activities in each fiscal year (based on a 40-hour work week); these reports are based on the hours that TNTP staff reported working on a given project each week. TNTP or staff within districts provided us with information about the staff who worked on TEACh program activities and the percentage of total time those district staff worked on these activities in each year.
- **Staff percentage effort on TEACh activities.** To allocate time of TNTP and district staff to specific components and activities, we provided contact person(s) in each district (typically TNTP staff) with an Excel tool that included our TEACh program activity list and asked them to provide information about the percentage of time that each individual spent on each TEACh program activity in each year. Contact person(s) generally gathered this information from each individual or were familiar enough with the individual staff members' work that they could provide the estimates themselves.
- **District salary data.** For each district employee that reportedly spent time on the TEACh program, as well as recruiting, hiring, and selection for new teachers more generally, we gathered individual salary data from public sources, as well as from the districts. Public data included names and salaries of most district employees. We shared these files with the district human resources offices, which verified or updated the information and provided salary information for anyone missing in the public data. Each district also provided each employee's hire and exit dates so we could derive the number of months an individual worked in a given school year to compute the relevant salary for each school year. Each district also gave us the average salary fringe rates for all staff or specific groups of staff, if rates were different, so we could apply the respective rate to the employee's base salary.
- **Interviews with TNTP and district staff.** Lastly, to both collect and verify our data, we conducted roughly two to three interviews per year in each district with TNTP staff and—on occasion—district personnel who managed or played key roles in the initiative to gather more information about costs from any expenditure or percentage time reports and gaps in costs.

Data Analysis

Annual cost estimates for each district—and average annual costs within and across districts—were calculated after determining both the nonpersonnel and total personnel costs in each year. Personnel costs were determined first by attributing a salary to each individual and then estimating that individual's time on each pipeline activity. The basic personnel cost is computed by multiplying salary by the proportion of an individual's time spent on the pipeline.

For part-time staff for whom we lacked salary data, we used their number of hours worked and the average hourly rate of their respective part-time position to compute their remuneration. To calculate personnel time on individual components of the TEACh pipeline (components outlined in Table A.5), we summed the total cost of all individuals' time spent on that component.

In addition to providing estimates for the cost of staff time, this report provides full-time equivalent (FTE) estimates overall and for each component. These data were calculated by first estimating the proportion of each full-time employee's time spent on each component (e.g., one employee is estimated to spend 20 percent of a full-time job, or 0.2 FTE, on component 1) and then summing everyone's time, or FTE, for each component and for the pipeline overall. For part-time staff, we divided their number of hours worked by 1,750 hours (40 hours per week times 52 weeks, taking into account typical vacation and sick time estimates) to estimate their FTE.

Lastly, we adjusted average annual estimates to account for regional variation in purchasing power (i.e., cost of living) and inflation. To make adjustments to account for regional variation in purchasing power, we identified regional price parities (RPPs)—which measure differences in price levels of goods and services—for each year for the metropolitan statistical area in which each district is located. We then divided the costs by the RPP adjustment factor (i.e., the RPP index divided by 100) for each year. To make adjustments for inflation, we used the Consumer Price Index for All Urban Consumers (CPI-U) to inflate expenditures from earlier years of the TEACh program to equivalent dollars for 2018.

Cost Study Methodological Limitations

There are challenges with collecting any self-reported data on costs and how staff allocated time. Specifically, the personnel data were gathered retrospectively, sometimes one to two years after activities had been undertaken. In some cases, we were unable to gather data on time spent on specific activities and could gather time estimates only at the component level. On the other hand, staff were reasonably confident in allocating time to specific activities at the component level, given that some staff job descriptions make it relatively easy to determine the components on which those staff spent their time (e.g., coaches typically spend time only on support and training components and not on hiring components). Thus, in this report, we summarize personnel time by component rather than by activity.

In addition, while these costs provide a useful overview of what it may take for districts to create these teacher pipelines, these figures may not capture all the costs that could be associated with these pipelines. For example, if the districts contracted with other consultants to support the TEACh work beyond what was supported through the SEED grant, those costs would not be captured. In addition, this report does not capture the costs of time for TEACh teachers, who spent considerable time attending trainings associated with their preparation and new-teacher support in their first program year, as well as forgoing any summer employment that they might have otherwise had. Instead, it focuses on the cost of the time for those running and supporting the TEACh program. In addition, some programs offered trainees stipends for their participation in the summer PST. Given the variation in how programs provided these stipends and the years in which they did so, we do not include them in the overall costs. However, we do report on them in a separate section of our cost report. In addition, our study may not adequately capture the costs for technology in which the districts may have invested to track and support TEACh teachers.

Lastly, TEACh program costs may not be completely comparable across districts. In particular, the District B TEACh program had been operating, albeit in a somewhat different form, before TNTP worked with District B to revise that program. Thus, District B costs may not be comparable to the costs for Districts A and C. In addition, while TEACh programs in all districts included the same basic components and activities, the programs emphasized these activities in different ways, depending on their context, and may have dealt with different challenges that impacted TEACh costs, which we don’t describe in detail in this report. In our reporting, we strive—when possible—to match cost findings with some implementation data that might help explain the cost differences we observed. Finally, these costs could be more informative if examined alongside costs for recruiting, hiring, and supporting first-year teachers in each district who did not participate in the TEACh program. However, it proved too difficult to gather data on these costs across multiple departments and personnel in the large urban districts in our study. Future studies could examine these costs to shed more light on the costs of first-year teacher recruitment, selection, hiring, and support regardless of teachers’ preparation routes.

References

Bang, H., and J. M. Robins, “Doubly Robust Estimation in Missing Data and Causal Inference Models,” *Biometrics*, Vol. 61, No. 4, 2005, pp. 962–973.

Benjamini, Y., and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 57, No. 1, 1995, pp. 289–300.

Cameron, A. Colin, and Douglas L. Miller, “A Practitioner’s Guide to Cluster-Robust Inference,” *Journal of Human Resources*, Vol. 50, No. 2, 2015, pp. 317–372.

Cameron, A. Colin, and Pravin K. Trivedi, *Microeconometrics: Methods and Applications,* New York: Cambridge University Press, 2005.

Chambers, Jay G., Irene Lam, and Kanya Mahitivanichcha, “Examining Context and Challenges in Measuring Investment in Professional Development: A Case Study of Six School Districts in the Southwest Region,” *Issues & Answers, Regional Education Laboratory Southwest*, No. 037, National Center for Education Evaluation and Regional Assistance, Institute of Educational Sciences, U.S. Department of Education, September 2008.

Cooper, Harris, Larry V. Hedges, and Jeffrey C. Valentine, eds., *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed., New York: Russell Sage Foundation, 2009.

Honaker, James, and Gary King, “What to Do About Missing Values in Time-Series Cross-Section Data,” *American Journal of Political Science*, Vol. 54, No. 2, 2010.

Kang, J. D., and J. L. Schafer, “Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data,” *Statistical Science*, Vol. 22, No. 4, 2007, pp. 523–539.

Kaufman, Julia H., Susan M. Gates, Melody Harvey, Yan Wang, and Mark Barrett, *What It Takes to Operate and Maintain Principal Pipelines: Costs and Other Resources*, Santa Monica, Calif.: RAND Corporation, RR-2078-WF, 2017. As of September 2, 2020: https://www.rand.org/pubs/research_reports/RR2078.html

Lee, B., J. Lessler, and E. Stuart, “Improving Propensity Score Weighting Using Machine Learning,” *Statistics in Medicine*, Vol. 29, 2010, pp. 337–346.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral, “Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies,” *Psychological Methods*, Vol. 9, No. 4, 2004, pp. 403–425.

Ridgeway, G., “Fast Doubly Robust Estimation,” R software package, 2014.

Ridgeway, G., D. Madigan, and T. Richardson, “Boosting Methodology for Regression Problems,” *Proceedings*, Seventh International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, Florida, January 3–6, 1999.

Rosenbaum, P. R., and D. B. Rubin, “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, Vol. 70, No. 1, 1983, pp. 41–55.

Schomaker, Michael, and Christian Heumann, “Bootstrap Inference When Using Multiple Imputation,” *Statistics in Medicine*, Vol. 37, No. 14, June 2018.

Appendix B. Impact Analysis Tables

In this section of the technical appendix, we include both results from additional exploratory analyses evaluating the impacts of the TEACh program, and tables describing the comparison groups of teachers and students considered in our impact analyses.

Table B.1. Characteristics of TEACh First-Year Teachers and Other First-Year Teachers, by District

	Overall		District A		District B		District C	
	TEACh Hires	Other New Hires	TEACh Hires	Other New Hires	TEACh Hires	Other New Hires	TEACh Hires	Other New Hires
Total N (% of total)	511 (13.2)	3,350 (86.8)	44 (4.5)	934 (95.5)	333 (15.3)	1,846 (84.7)	134 (19.0)	570 (81.0)
Bilingual certified (%)	25.4	12.1	43.2	17.0	30.3	11.9	7.5	4.6
Special education certified (%)	14.9	12.9	27.3	28.6	9.3	6.9	24.6	6.7
Emergency certified (%)	n/a	n/a	n/a	n/a	n/a	n/a	11.9	18.6
Taught mathematics (%)	49.7	32.3	25.0	20.7	48.0	33.3	61.9	47.9
Taught in tested grades and subjects (%)	n/a	n/a	n/a	n/a	31.2	28.1	41.0	27.0
Female (%)	n/a	n/a	n/a	n/a	67.2	71.7	68.7	67.4
White (%)	33.3	45.5	52.5	55.9	33.0	41.4	27.6	41.9
Black (%)	20.8	21.5	18.2	21.6	23.5	26.0	14.9	7.0
Hispanic (%)	33.0	20.7	6.8	12.0	35.9	25.0	34.3	21.2
Asian (%)	7.8	7.4	9.1	6.8	5.1	3.6	14.2	20.7
Other / unknown (%)	5.1	4.8	13.7	3.7	2.4	4.0	9.0	9.1

NOTE: *Taught mathematics (%)* was defined differently in each district, depending on how courses are labeled and how teachers are linked to courses in administrative records. Data from District A span new hires from school year 2016–2017 and 2017–2018. District B data span 2016–2017 through 2018–2019. District C data span 2017–2018 and 2018–2019.

Table B.2. Student and School Characteristics of TEACh First-Year Teachers and Other First-Year Teachers, by District

	District A		District B		District C	
	TEACh Hires	Other New Hires	TEACh Hires	Other New Hires	TEACh Hires	Other New Hires
Teachers' students						
Special education (%)	n/a	n/a	9.3	9.5	49.6	20.1
ELL (%)	n/a	n/a	46.8	41.3	38.2	26.8
FRPL (%)	n/a	n/a	68.3	70.3	n/a	n/a
White (%)	n/a	n/a	2.4	3.6	6.9	11.0
Black (%)	n/a	n/a	23.9	21.7	16.3	11.0
Hispanic (%)	n/a	n/a	65.7	68.3	47.2	35.2
Asian (%)	n/a	n/a	n/a	n/a	13.3	23.7
Other/unknown race (%)	n/a	n/a	2.0	0.9	16.2	19.1
Average standardized prior-year ELA scores	n/a	n/a	–0.101	–0.066	–0.977	–0.376
Average standardized prior-year mathematics scores	n/a	n/a	–0.185	–0.046	–0.990	–0.378
Teachers' schools						
Special education (%)	21.3	22.1	7.9	7.9	16.8	15.5
ELL (%)	37.6	33.1	45.8	44.0	34.0	28.0
FRPL (%)	81.3	78.7	71.4	70.9	n/a	n/a
White (%)	5.9	8.7	3.6	4.1	10.6	11.4
Black (%)	38.7	39.7	23.8	20.9	8.4	8.8
Hispanic (%)	50.2	44.0	66.1	69.1	42.4	34.9
Asian (%)	4.3	6.4	n/a	n/a	19.2	25.1
Other/unknown race (%)	0.9	1.1	6.5	5.9	19.4	19.8
Average standardized prior-year ELA scores	–0.244	–0.115	–0.078	–0.053	–0.297	–0.217
Average standardized prior-year mathematics scores	–0.253	–0.127	–0.086	–0.065	–0.335	–0.246

NOTE: FRPL = free or reduced-price lunch. Student test scores are standardized relative to their districtwide sample in the same grade.

Table B.3. TEACh Student Achievement Gains Relative to Students of Comparison Second-Year Teachers, Overall and by District

	Multidistrict Average		District B		District C	
	ELA	Mathematics	ELA	Mathematics	ELA	Mathematics
TEACh	0.026	0.061	0.031	0.073	0.007	0.044
Confidence interval (95%)	[–0.060 to 0.111]	[–0.101 to 0.224]	[–0.068 to 0.130]	[–0.143 to 0.290]	[–0.177 to 0.192]	[–0.215 to 0.303]
N of TEACh teachers	n/a		27	29	35	35
N of comparison teachers	n/a		175	130	178	179
N of TEACh students	n/a		790	1059	266	189
Effective N of comparison students (post-weighting)	n/a		2,745	1,193	235	99
Cohorts	n/a		Cohorts 1 and 2		Cohort 1	
School years	n/a		2017–2018 and 2018–2019		2018–2019	

NOTE: Effective Ns correspond to the number of comparison observations times their fractional weight given in our analysis.

Table B.4. TEACh Student Achievement Gains Relative to Students of Comparison First-Year Teachers, Overall and by District

	Multidistrict Average		District B		District C	
	ELA	Mathematics	ELA	Mathematics	ELA	Mathematics
TEACh	0.035	0.080*	0.053~	0.065	–0.002	0.101~
Confidence interval (95%)	[–0.011 to 0.080]	[0.004 to 0.156]	[–0.003 to 0.109]	[–0.036 to 0.165]	[–0.082 to 0.079]	[–0.020 to 0.222]
N of TEACh teachers	n/a		62	62	81	81
N of comparison teachers	n/a		326	243	413	406
N of TEACh students	n/a		2,297	2,627	809	565
Effective N of comparison students (post-weighting)	n/a		8,597	4,110	239	609
Cohorts	n/a		Cohorts 1, 2, and 3		Cohorts 1 and 2	
School years	n/a		2016–2017 through 2018–2019		2017–2018 and 2018–2019	

NOTE: Effective Ns correspond to the number of comparison observation times their fractional weight given in our analysis.

Table B.5. TEACh District Evaluation Scores Relative to Comparison Second-Year Teachers

	District A	District B	District C
Evaluation scale range	1 to 4	1 to 100	1 to 5
TEACh	–0.079~	–0.398	–0.026
Confidence interval	[–0.162 to 0.003]	[–2.470 to 1.680]	[–0.455 to 0.403]
N of TEACh teachers	10	163	34
Effective N of comparison teachers (post-weighting)	22	476	45
Cohorts	Cohort 1	Cohorts 1 and 2	Cohort 1
School years	2017–2018	2017–2018 and 2018–2019	2018–2019

NOTE: Effective Ns correspond to the number of comparison observations times their fractional weight given in our analysis.

Table B.6. TEACh District Evaluation Scores Relative to Comparison First-Year Teachers

	District A	District B	District C
Evaluation scale range	1 to 4	1 to 100	1 to 5
TEACh	–0.017	–1.550*	–0.191
Confidence interval	[–0.163 to 0.129]	[–3.020 to –0.082]	[–0.468 to 0.086]
N of TEACh teachers	36	306	71
Effective N of comparison teachers (post-weighting)	272	1,168	67
Cohorts	Cohort 1 and 2	Cohorts 1, 2, and 3	Cohorts 1 and 2
School years	2016–2017 and 2017–2018	2016–2017 through 2018–2019	2017–2018 and 2018–2019

NOTE: Effective Ns correspond to the number of comparison observation times their fractional weight given in our analysis.

Table B.7. Within-District Retention Rates of Teach Teachers Between Their Second and Third Year, Overall and for Individual Districts

	Multidistrict Average	District A	District B	District C
TEACh	0.008	0.060	–0.017	0.025
Confidence interval (95%)	[–0.056 to 0.072]	[–0.074 to 0.194]	[–0.100 to 0.066]	[–0.138 to 0.189]
N of TEACh teachers	n/a	15	173	51
Effective N of comparison teachers (post-weighting)	n/a	55	532	48
Cohorts	n/a	Cohort 1	Cohorts 1 and 2	Cohort 1
School years	n/a	2017–2018	2017–2018 and 2018–2019	2018–2019

NOTE: Effective Ns correspond to the number of comparison observations times their fractional weight given in our analysis.

Table B.8. Within-District Retention Rates of Teach Teachers Between Their First and Second Year, Overall and for Individual Districts

	Multidistrict Average	District A	District B	District C
TEACh	–0.025	–0.141~	–0.019	0.004
Confidence interval (95%)	[–0.067 to 0.018]	[–0.283 to 0.001]	[–0.073 to 0.034]	[–0.081 to 0.088]
N of TEACh teachers	n/a	44	333	134
Effective N of comparison teachers (post-weighting)	n/a	637	1,270	140
Cohorts	n/a	Cohorts 1 and 2	Cohorts 1, 2, and 3	Cohorts 1 and 2
School years	n/a	2016–2017 and 2017–2018	2016–2017 through 2018–2019	2017–2018 and 2018–2019

NOTE: Effective Ns correspond to the number of comparison observations times their fractional weight given in our analysis.

Table B.9. Estimated Coefficients, Confidence Intervals, and Teach Sample Sizes for All Outcomes, by Cohort and Year, for Each District in the Study

	ELA	Mathematics	Evaluations	Retention
RESULTS KEY	*TEACh Coefficient* *[95% Confidence Interval]* *N of TEACh Students*	*TEACh Coefficient* *[95% Confidence Interval]* *N of TEACh Students*	*TEACh Coefficient* *[95% Confidence Interval]* *N of TEACh Teachers*	*TEACh Coefficient* *[95% Confidence Interval]* *N of TEACh Teachers*
District A				
Cohort 1 year 1	n/a	n/a	–0.040 [–0.292 to 0.213] 14	–0.066 [–0.271 to 0.139] 22
Cohort 1 year 2	n/a	n/a	–0.079~ [–0.162 to 0.003] 10	0.060 [–0.074 to 0.194] 15
Cohort 2 year 1	n/a	n/a	–0.008 [–0.238 to 0.221] 22	–0.197 [–0.490 to 0.096] 22
District B				
Cohort 1 year 1	0.044 [–0.071 to 0.159] 718	0.122 [–0.045 to 0.289] 1331	–2.720* [–5.240 to -0.204] 115	0.047 [–0.034 to 0.128] 122
Cohort 1 year 2	0.159* [0.0121 to 0.306] 325	0.089 [–0.255 to 0.433] 728	–1.680 [–4.480 to 1.120] 93	–0.047 [–0.180 to 0.087] 94
Cohort 1 year 3	0.105 [–0.086 to 0.296] 359	0.195 [–0.095 to 0.485] 464	3.200~ [–0.149 to 6.550] 64	0.031 [–0.069 to 0.131] 66

	ELA	Mathematics	Evaluations	Retention
Cohort 2 year 1	0.071~ [–0.002 to 0.145] 566	–0.009 [–0.189 to 0.170] 406	0.602 [–2.160 to 3.370] 93	–0.060 [–0.178 to 0.057] 109
Cohort 2 year 2	–0.052 [–0.177 to 0.073] 465	0.070 [–0.155 to 0.295] 331	0.924 [–2.430 to 4.270] 70	0.043 [–0.068 to 0.153] 79
Cohort 3 year 1	0.041 [–0.056 to 0.138] 1013	–0.004 [–0.147 to 0.139] 890	–1.050 [–4.220 to 2.120] 98	0.004 [–0.081 to 0.089] 102
District C				
Cohort 1 year 1	0.007 [–0.128 to 0.143] 362	0.164* [0.008 to 0.320] 271	–0.226 [–0.719 to 0.267] 29	–0.069 [–0.226 to 0.087] 64
Cohort 1 year 2	0.007 [–0.177 to 0.192] 266	0.044 [–0.215 to 0.303] 189	–0.026 [–0.455 to 0.403] 34	0.025 [–0.138 to 0.189] 51
Cohort 1 year 2 second-year evaluations only	n/a	n/a	–0.189 [–0.662 to 0.284] 28	n/a
Cohort 2 year 1	–0.027 [–0.213 to 0.159] 447	0.091 [–0.115 to 0.297] 294	–0.270 [–0.596 to 0.056] 42	0.088 [–0.031 to 0.206] 70

NOTE: Cohort 1 year 2 *second-year evaluations only* refers to estimates for a sample just of second-year teachers who received their evaluations in year 2. ~ = p < 0.1. * = p < 0.05. ** = p < 0.01. *** = p < 0.001. None of the nominally significant results shown were robust to adjustment for multiple (cohort and/or district) comparisons when applying a false discovery threshold of 10 percent. Effective Ns correspond to the number of comparison observations times their fractional weight given in our analysis.

Table B.10. First Year Teachers' Evaluation Ratings in District B, Overall and for Teachers Included in Student Achievement Analyses

	Overall first-year teachers	Teachers of ELA in Grades 4–8	Teachers of mathematics in Grades 4–8
Evaluation scale	1 to 100	1 to 100	1 to 100
TEACh	–1.550*	0.124	–2.370
Confidence interval (95%)	[–3.020 to –0.082]	[–3.320 to 3.570]	[–6.410 to 1.670]
N of TEACh teachers	306	61	60
Effective N of comparison teachers (post-weighting)	1,168.0	210.2	130.7

NOTE: Effective Ns correspond to the number of comparison observations times their fractional weight given in our analysis.

Table B.11. Average Levels of Observable Characteristics in First-Year Teacher Retention Outcome Analyses

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
District A						
School average test score	–0.12	–0.21	–0.24	0.26	0.10	–0.09
Percentage of school on free/reduced lunch	0.79	0.80	0.81	0.17	0.12	0.13
English language learner	0.17	0.37	0.43	0.26	0.06	0.16
Special needs	0.29	0.25	0.27	–0.01	0.02	0.06
Primary grade levels						
Elementary school	0.28	0.26	0.39	0.11	0.13	0.35
Middle school	0.36	0.33	0.30	–0.07	–0.04	–0.10
High school	0.36	0.41	0.32	–0.04	–0.09	–0.23
2017–2018 school year	0.50	0.47	0.50	0.00	0.03	0.06
District B						
Average student characteristics						
Prior ELA score[a]	–0.25	–0.26	–0.26	0.07	0.06	0.01
Prior math score[a]	–0.20	–0.24	–0.29	0.16	0.10	–0.10
Black/Hispanic	0.90	0.90	0.89	0.05	0.03	–0.02
Special needs	0.11	0.13	0.13	0.07	0.04	0.02
Free/reduced lunch	0.70	0.68	0.68	0.09	0.03	0.00
At risk	0.56	0.58	0.59	0.08	0.04	0.05
Disciplinary actions	0.33	0.32	0.31	0.16	0.05	–0.02
Absence percentage	0.05	0.05	0.05	0.05	0.03	0.00
English language learner	0.40	0.45	0.46	0.16	0.05	0.04
Other characteristics						
Evaluation category 1	0.52	0.49	0.47	–0.04	–0.02	–0.04
Evaluation category 2	0.13	0.19	0.21	0.08	0.03	0.10
Evaluation category 3	0.17	0.09	0.06	–0.11	–0.02	–0.20
Evaluation category 4	0.18	0.23	0.25	0.07	0.01	0.05
School year						
2016–2017	0.33	0.34	0.37	0.04	0.03	0.07
2017–2018	0.37	0.35	0.33	–0.04	–0.02	–0.05
2018-–019	0.30	0.31	0.31	0.00	–0.01	–0.02
Primary grade level	5.80	5.02	4.88	0.17	0.05	–0.04
School average test score	–0.06	–0.06	–0.08	0.08	0.08	–0.09
Percentage of school on free/reduced lunch	0.71	0.70	0.71	0.05	0.07	0.09

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
District C						
Average student characteristics						
Prior ELA score[a]	–0.36	–0.84	–0.90	0.36	0.08	–0.10
Prior math score[a]	–0.36	–0.88	–0.91	0.38	0.09	–0.04
Black/Hispanic	0.46	0.59	0.63	0.31	0.08	0.15
Absence percentage	0.05	0.06	0.06	0.17	0.06	0.00
Special needs	0.20	0.46	0.50	0.38	0.09	0.11
English language learner	0.27	0.38	0.39	0.28	0.05	0.04
Disciplinary actions	0.06	0.09	0.12	0.18	0.10	0.09
Primary grade levels						
Elementary school	0.35	0.49	0.48	0.13	–0.01	–0.03
Middle school	0.34	0.17	0.19	–0.15	0.02	0.10
High school	0.30	0.34	0.33	0.02	–0.01	–0.03
English language learner specialist	0.37	0.46	0.44	0.07	–0.02	–0.06
Special needs specialist						
Mild to moderate	0.05	0.35	0.44	0.39	0.09	0.24
Moderate to severe	0.02	0.04	0.04	0.02	0.00	0.03
None	0.93	0.61	0.51	–0.41	–0.09	–0.23
2018–2019 school year	0.51	0.52	0.52	0.01	0.00	–0.01
School average test score	–0.23	–0.31	–0.31	0.15	0.10	–0.01
Percentage of school that is Black/Hispanic	0.44	0.49	0.51	0.16	0.08	0.06

NOTES: This table depicts the covariate balance in one of the five multiple imputation runs of the analysis, prior to the regression stage. The Kolmogorov-Smirnov (KS) statistic is a nonparametric measure of similarity between the distribution of two variables. The *Adjusted Difference* column is the difference between the group average values in standard deviation units, based on Hedges' G statistic, except that we adapted the formula for the control group to include the weighted standard deviation and the effective sample size after weighting.

[a] Included only in subgroup comparisons that focused on teachers linked to students in the student achievement analyses.

Table B.12. Average Levels of Observable Characteristics in First-Year ELA Student Test Score Analyses

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
District B						
Prior ELA score	–0.12	–0.13	–0.10	0.02	0.02	0.04
Prior math score	–0.06	–0.05	–0.01	0.03	0.02	0.04
Race/Ethnicity/Gender						
White	0.04	0.04	0.06	0.02	0.01	0.18
Black	0.22	0.20	0.19	–0.04	–0.01	–0.05
Hispanic	0.72	0.73	0.73	0.01	0.00	–0.01

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
Asian	0.01	0.01	0.02	0.01	0.00	0.18
Other	0.01	0.01	0.01	0.00	0.00	–0.06
Female	0.47	0.47	0.46	0.00	–0.01	–0.02
Other Characteristics						
Disciplinary actions	0.34	0.43	0.45	0.14	0.02	0.02
Prior absence percentage	0.03	0.03	0.03	0.04	0.01	–0.02
Special needs	0.08	0.13	0.13	0.04	0.00	0.00
At risk	0.70	0.66	0.64	–0.06	–0.01	–0.04
Free/reduced lunch	0.79	0.74	0.71	–0.08	–0.02	–0.07
English language learner	0.48	0.49	0.49	0.01	0.00	0.00
School year						
2016–2017	0.36	0.32	0.31	–0.05	–0.01	–0.02
2017–2018	0.39	0.26	0.25	–0.15	–0.01	–0.05
2018–2019	0.25	0.42	0.44	0.19	0.02	0.06
Grade level	6.36	5.96	5.96	0.16	0.02	0.00
School average test score	–0.03	–0.02	0.03	0.14	0.17	0.12
Percentage of school on free/reduced lunch	0.76	0.70	0.68	0.26	0.12	–0.11
District C						
Prior ELA score	–0.18	–0.50	–0.41	0.12	0.04	0.09
Prior math score	–0.17	–0.48	–0.45	0.12	0.03	0.03
Race/Ethnicity/Gender						
White	0.12	0.10	0.11	–0.01	0.02	0.10
Black	0.08	0.10	0.11	0.03	0.02	0.12
Hispanic	0.33	0.42	0.42	0.09	0.00	–0.01
Asian	0.35	0.26	0.23	–0.12	–0.03	–0.10
Other	0.11	0.12	0.12	0.01	0.00	0.00
Female	0.48	0.46	0.45	–0.03	–0.01	–0.03
Other Characteristics						
Disciplinary actions	0.02	0.02	0.04	0.02	0.03	0.09
Prior absence percentage	0.03	0.04	0.03	0.09	0.04	–0.08
English language learner	0.23	0.31	0.31	0.08	0.00	–0.01
Mother's education level						
Less than high school	0.06	0.05	0.04	–0.03	–0.01	–0.18
High school diploma	0.07	0.06	0.05	–0.02	0.00	–0.04
Some college	0.05	0.02	0.02	–0.02	0.00	0.03
Bachelor's degree	0.05	0.02	0.02	–0.03	0.00	–0.01

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
Graduate degree	0.03	0.02	0.03	0.00	0.01	0.26
Declined to answer	0.74	0.83	0.83	0.10	0.00	0.02
Linked to special needs specialist						
Mild to moderate	0.10	0.22	0.31	0.21	0.09	0.30
Moderate to severe	0.01	0.01	0.01	–0.01	–0.01	–0.33
None	0.89	0.77	0.68	–0.21	–0.09	–0.27
Linked to English language learner specialist	0.67	0.72	0.67	–0.01	–0.05	–0.15
Other students linked to associated teacher						
English language learner	0.24	0.31	0.31	0.22	0.09	–0.01
Non–special needs only	0.86	0.75	0.76	–0.10	0.01	0.04
Some special needs	0.14	0.18	0.15	0.01	–0.03	–0.12
All special needs	0.00	0.07	0.09	0.09	0.01	0.12
2018–2019 school year	0.44	0.54	0.55	0.11	0.01	0.03
Grade level	5.98	5.21	5.24	0.30	0.03	0.02
School average test score	–0.17	–0.26	–0.27	0.20	0.11	–0.02
Percentage of school that is Black/Hispanic	0.41	0.46	0.47	0.20	0.07	0.03

NOTES: This table depicts the covariate balance in one of the five multiple imputation runs of the analysis, prior to the regression stage. The Kolmogorov-Smirnov (KS) statistic is a nonparametric measure of similarity between the distribution of two variables. The *Adjusted Difference* column is the difference between the group average values in standard deviation units, based on Hedges' G statistic, except that we adapted the formula for the control group to include the weighted standard deviation and the effective sample size after weighting.

Table B.13. Average Levels of Observable Characteristics in First-Year Math Student Test Score Analyses

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
District B						
Prior ELA score	–0.11	–0.12	–0.11	0.02	0.01	0.01
Prior math score	–0.07	–0.16	–0.16	0.04	0.01	0.00
Race/Ethnicity/Gender						
White	0.04	0.04	0.05	0.01	0.01	0.08
Black	0.21	0.20	0.19	–0.02	–0.01	–0.03
Hispanic	0.73	0.73	0.74	0.00	0.00	0.01
Asian	0.01	0.01	0.02	0.01	0.00	0.05
Other	0.01	0.01	0.01	0.00	0.00	0.03
Female	0.47	0.46	0.50	0.03	0.03	0.08

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
Other characteristics						
Disciplinary actions	0.37	0.35	0.34	0.02	0.01	–0.01
Prior absence percentage	0.03	0.03	0.03	0.04	0.02	–0.01
Special needs	0.09	0.09	0.09	0.00	0.00	–0.01
At risk	0.72	0.73	0.73	0.01	–0.01	–0.02
Free/reduced lunch	0.77	0.76	0.76	–0.01	0.00	0.00
English language learner	0.49	0.55	0.55	0.06	0.00	–0.01
School year						
2016–2017	0.27	0.49	0.51	0.24	0.01	0.03
2017–2018	0.45	0.16	0.15	–0.30	–0.01	–0.04
2018–2019	0.28	0.34	0.34	0.06	0.00	–0.01
Grade level	6.31	5.72	5.74	0.28	0.02	0.01
School average test score	–0.04	–0.03	–0.07	0.08	0.12	–0.12
Percentage of school on free/reduced lunch	0.76	0.74	0.75	0.09	0.12	0.06
District C						
Prior ELA score	–0.16	–0.48	–0.44	0.13	0.03	0.04
Prior math score	–0.17	–0.51	–0.52	0.15	0.03	–0.01
Race/Ethnicity/Gender						
White	0.12	0.08	0.09	–0.03	0.01	0.10
Black	0.09	0.11	0.11	0.02	0.01	0.03
Hispanic	0.33	0.45	0.47	0.15	0.03	0.06
Asian	0.34	0.23	0.19	–0.15	–0.04	–0.15
Other	0.12	0.14	0.13	0.02	0.00	–0.02
Female	0.48	0.46	0.47	–0.01	0.01	0.02
Other Characteristics						
Disciplinary actions	0.02	0.03	0.05	0.01	0.02	0.06
Prior absence percentage	0.03	0.04	0.04	0.11	0.05	–0.07
English language learner	0.21	0.31	0.32	0.11	0.01	0.02
Mother's education level						
Less than high school	0.07	0.04	0.04	–0.03	–0.01	–0.12
High school diploma	0.08	0.06	0.05	–0.03	–0.01	–0.08
Some college	0.06	0.02	0.03	–0.02	0.01	0.25
Bachelor's degree	0.07	0.02	0.02	–0.05	0.00	–0.07
Graduate degree	0.03	0.01	0.02	–0.01	0.01	0.49
Declined to answer	0.70	0.85	0.84	0.14	–0.01	–0.04
Linked to special needs specialist						

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
Mild to moderate	0.07	0.13	0.15	0.07	0.02	0.10
Moderate to severe	0.00	0.02	0.01	0.01	–0.01	–0.21
None	0.92	0.85	0.84	–0.08	–0.01	–0.07
Linked to English language learner specialist	0.67	0.77	0.79	0.12	0.02	0.06
Other students linked to associated teacher						
English language learner	0.21	0.31	0.33	0.36	0.06	0.09
Non–special needs only	0.85	0.75	0.75	–0.10	0.00	–0.01
Some special needs	0.14	0.14	0.13	–0.01	–0.01	–0.03
All special needs	0.01	0.11	0.12	0.11	0.01	0.05
2018–2019 school year	0.37	0.54	0.52	0.15	–0.02	–0.04
Grade level	5.93	4.90	4.82	0.36	0.03	–0.08
School average test score	–0.18	–0.31	–0.35	0.25	0.16	–0.09
Percentage of school that is Black/Hispanic	0.41	0.49	0.52	0.26	0.13	0.11

NOTES: This table depicts the covariate balance in one of the five multiple imputation runs of the analysis, prior to the regression stage. The Kolmogorov-Smirnov (KS) statistic is a nonparametric measure of similarity between the distribution of two variables. The Adjusted Difference column is the difference between the group average values in standard deviation units, based on Hedges' G statistic, except that we adapted the formula for the control group to include the weighted standard deviation and the effective sample size after weighting.

Table B.14. Average Levels of Observable Characteristics in Second-Year Teacher Retention Outcome Analyses

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
District A						
School average test score	–0.17	–0.09	–0.09	0.26	0.05	–0.01
Percentage of school on free/reduced lunch	0.78	0.79	0.78	0.17	0.07	–0.02
English language learner	0.18	0.38	0.40	0.22	0.02	0.06
Special needs	0.29	0.11	0.07	–0.22	–0.04	–0.31
Primary grade levels						
Elementary school	0.30	0.37	0.40	0.10	0.03	0.09
Middle school	0.32	0.23	0.20	–0.12	–0.03	–0.10
High school	0.38	0.40	0.40	0.02	0.00	–0.01
District B						
Average student characteristics						

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
Prior ELA score[a]	–0.20	–0.26	–0.28	0.14	0.14	–0.04
Prior math score[a]	–0.17	–0.29	–0.23	0.16	0.10	0.12
Black/Hispanic	0.89	0.91	0.91	0.13	0.06	0.07
Special needs	0.12	0.13	0.14	0.15	0.07	0.03
Free/reduced lunch	0.61	0.61	0.61	0.07	0.05	0.01
At risk	0.53	0.55	0.57	0.14	0.07	0.08
Disciplinary actions	0.51	0.44	0.43	0.16	0.05	–0.01
Absence percentage	0.05	0.05	0.05	0.06	0.04	0.02
English language learner	0.43	0.48	0.50	0.20	0.07	0.06
Other characteristics						
2018–2019 school year	0.52	0.47	0.46	–0.07	–0.01	–0.03
Evaluation category 1	0.58	0.50	0.46	–0.12	–0.04	–0.09
Evaluation category 2	0.13	0.20	0.20	0.07	0.00	0.01
Evaluation category 3	0.14	0.07	0.05	–0.09	–0.02	–0.23
Evaluation category 4	0.14	0.24	0.29	0.14	0.05	0.17
Primary grade level	6.17	4.75	4.37	0.26	0.06	–0.10
School average test score	–0.05	–0.05	–0.08	0.10	0.10	–0.12
Percentage of school on free/reduced lunch	0.62	0.64	0.65	0.14	0.09	0.07
District C						
Average student characteristics						
Prior ELA score[a]	–0.28	–0.88	–1.09	0.51	0.18	–0.37
Prior math score[a]	–0.27	–0.89	–1.06	0.48	0.25	–0.27
Black/Hispanic	0.44	0.58	0.62	0.35	0.15	0.15
Absence percentage	0.04	0.04	0.05	0.24	0.19	0.28
Special needs	0.18	0.46	0.55	0.48	0.16	0.22
English language learner	0.29	0.40	0.38	0.34	0.09	–0.10
Disciplinary actions	0.05	0.10	0.12	0.23	0.05	0.06
Primary grade levels						
Elementary school	0.38	0.53	0.59	0.21	0.06	0.14
Middle school	0.29	0.26	0.27	–0.01	0.01	0.03
High school	0.34	0.20	0.14	–0.20	–0.07	–0.29
English language learner specialist	0.42	0.49	0.47	0.05	–0.02	–0.04
Special needs specialist						
Mild to moderate	0.04	0.33	0.43	0.39	0.10	0.25
Moderate to severe	0.01	0.04	0.02	0.01	–0.02	–0.41
None	0.95	0.63	0.55	–0.40	–0.08	–0.20

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
School average test score	–0.22	–0.22	–0.31	0.20	0.18	–0.20
Percentage of school that is Black/Hispanic	0.43	0.47	0.51	0.23	0.17	0.17
Year of most recent evaluation[b]	0.83	0.79	0.82	–0.01	0.03	0.12

NOTES: This table depicts the covariate balance in one of the five multiple imputation runs of the analysis, prior to the regression stage. The Kolmogorov-Smirnov (KS) statistic is a nonparametric measure of similarity between the distribution of two variables. The Adjusted Difference column is the difference between the group average values in standard deviation units, based on Hedges' G statistic, except that we adapted the formula for the control group to include the weighted standard deviation and the effective sample size after weighting.

[a] Included only in subgroup comparisons that focused on teachers linked to students in the student achievement analyses.

[b] Included only in second-year comparison of most recent evaluation.

Table B.15. Average Levels of Observable Characteristics in Second-Year ELA Student Test Score Analyses

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
District B						
Prior ELA score	–0.06	–0.09	–0.10	0.05	0.01	0.00
Prior math score	–0.02	–0.10	–0.12	0.07	0.02	–0.02
Race/Ethnicity/Gender						
White	0.03	0.04	0.04	0.01	0.01	0.10
Black	0.16	0.17	0.17	0.02	0.00	0.01
Hispanic	0.79	0.77	0.76	–0.03	–0.01	–0.03
Asian	0.01	0.01	0.02	0.00	0.00	0.15
Other	0.01	0.01	0.01	0.00	0.00	–0.19
Female	0.46	0.46	0.46	0.00	0.00	0.01
Other Characteristics						
Disciplinary actions	0.48	0.57	0.57	0.13	0.01	0.00
Prior absence percentage	0.03	0.03	0.03	0.04	0.01	–0.02
Special needs	0.08	0.18	0.18	0.11	0.00	0.01
At risk	0.70	0.61	0.60	–0.10	0.00	–0.01
Free/reduced lunch	0.73	0.67	0.66	–0.07	–0.01	–0.04
English language learner	0.54	0.48	0.47	–0.07	–0.01	–0.03
2018–2019 school year	0.42	0.57	0.59	0.17	0.02	0.04
Grade level	6.34	5.35	5.34	0.38	0.02	–0.01
School average test score	–0.03	–0.01	–0.02	0.17	0.15	–0.03
Percentage of school on free/reduced lunch	0.69	0.64	0.64	0.27	0.19	0.02
District C						
Prior ELA score	0.02	-0.42	-0.41	0.21	0.05	0.01
Prior math score	0.05	-0.36	-0.40	0.21	0.05	-0.04
Race/Ethnicity/Gender						
White	0.14	0.07	0.08	-0.06	0.01	0.06
Black	0.07	0.11	0.09	0.02	-0.02	-0.12
Hispanic	0.24	0.31	0.42	0.18	0.11	0.29
Asian	0.43	0.38	0.27	-0.16	-0.11	-0.30
Other	0.13	0.12	0.14	0.01	0.01	0.07
Female	0.49	0.52	0.52	0.04	0.00	0.00
Other Characteristics						
Disciplinary	0.01	0.01	0.00	0.01	0.02	–0.13

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
actions						
Prior absence percentage	0.01	0.03	0.03	0.16	0.08	0.09
English language learner	0.21	0.35	0.34	0.13	–0.01	–0.04
Mother's education level						
Less than high school	0.03	0.03	0.04	0.00	0.00	0.07
high school diploma	0.05	0.04	0.05	–0.01	0.00	0.05
Some college	0.04	0.07	0.03	–0.01	–0.04	–0.52
Bachelor's degree	0.05	0.02	0.05	0.00	0.03	0.63
Graduate degree	0.03	0.01	0.03	0.01	0.02	0.53
Declined to answer	0.79	0.82	0.80	0.02	–0.02	–0.08
Linked to mild/moderate special education specialist	0.15	0.20	0.13	–0.03	–0.08	–0.34
Linked to English language learner specialist	0.70	0.84	0.88	0.19	0.04	0.21
Other students linked to associated teacher						
English language learner	0.21	0.34	0.36	0.61	0.12	0.08
Non–special needs only	0.86	0.82	0.79	–0.07	–0.03	–0.11
Some special needs	0.14	0.16	0.18	0.05	0.02	0.07
All special needs	0.00	0.02	0.03	0.03	0.01	0.27
Grade level	5.77	5.37	5.03	0.33	0.13	–0.30
School average test score	0.06	–0.16	–0.36	0.55	0.49	–0.62
Percentage of school that is Black/Hispanic	0.30	0.40	0.55	0.59	0.37	0.68

NOTES: This table depicts the covariate balance in one of the five multiple imputation runs of the analysis, prior to the regression stage. The Kolmogorov-Smirnov (KS) statistic is a nonparametric measure of similarity between the distribution of two variables. The *Adjusted Difference* column is the difference between the group average values in standard deviation units, based on Hedges' G statistic, except that we adapted the formula for the control group to include the weighted standard deviation and the effective sample size after weighting.

Table B.16. Average Levels of Observable Characteristics in Second-Year Math Student Test Score Analyses

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
District B						
Prior ELA score	–0.07	–0.09	–0.07	0.03	0.03	0.02
Prior math score	–0.03	–0.11	–0.10	0.06	0.02	0.01
Race/Ethnicity/Gender						
White	0.04	0.04	0.05	0.00	0.01	0.08
Black	0.20	0.18	0.17	–0.02	–0.01	–0.04
Hispanic	0.75	0.76	0.77	0.02	0.00	0.01
Asian	0.00	0.00	0.00	0.00	0.00	–0.29
Other	0.01	0.01	0.01	0.01	0.00	0.20
Female	0.48	0.51	0.55	0.07	0.04	0.10
Other Characteristics						
Disciplinary actions	0.63	0.36	0.33	0.29	0.03	–0.05
Prior absence percentage	0.03	0.03	0.03	0.04	0.02	0.03
Special needs	0.09	0.15	0.13	0.05	–0.02	–0.08
At risk	0.63	0.73	0.72	0.09	–0.01	–0.02
Free/reduced lunch	0.65	0.77	0.78	0.13	0.01	0.03
English language learner	0.50	0.52	0.50	0.01	–0.02	–0.05
2018-2019 school year	0.63	0.35	0.31	–0.31	–0.04	–0.11
Grade level	6.40	5.48	5.50	0.47	0.02	0.03
School average test score	0.00	0.01	–0.06	0.17	0.15	–0.22
Percentage of school on free/reduced lunch	0.63	0.69	0.72	0.39	0.19	0.28
District C						
Prior ELA score	–0.10	–0.40	–0.51	0.21	0.08	–0.12
Prior math score	–0.10	–0.31	–0.44	0.16	0.09	–0.14
Race/Ethnicity/Gender						
White	0.13	0.04	0.04	–0.09	0.00	0.04
Black	0.08	0.06	0.10	0.02	0.04	0.33
Hispanic	0.30	0.28	0.39	0.09	0.11	0.30
Asian	0.36	0.46	0.33	–0.03	–0.13	–0.33
Other	0.13	0.15	0.13	0.00	–0.02	–0.11
Female	0.50	0.52	0.50	0.01	–0.02	–0.04
Other Characteristics						
Disciplinary actions	0.02	0.01	0.00	0.02	0.01	–0.13
Prior absence percentage	0.02	0.02	0.04	0.14	0.15	0.23
English language	0.25	0.42	0.41	0.17	0.00	–0.01

	Unadjusted Non-TEACh Group	Adjusted Non-TEACh Group	TEACh Group	KS Before Weighting	KS After Weighting	Adjusted Difference (SD)
learner						
Mother's education level						
Less than high school	0.04	0.05	0.04	0.00	–0.01	–0.12
High school diploma	0.05	0.04	0.06	0.01	0.02	0.30
Some college	0.04	0.04	0.04	0.00	0.00	0.04
Bachelor’s degree	0.03	0.01	0.03	0.00	0.02	0.52
Graduate degree	0.03	0.01	0.02	–0.01	0.01	0.42
Declined to answer	0.82	0.85	0.81	–0.01	–0.04	–0.18
Linked to mild/moderate special education specialist	0.05	0.07	0.07	0.03	0.00	0.03
Linked to English language learner specialist	0.71	0.90	0.84	0.12	–0.07	–0.37
Other students linked to associated teacher						
English language learner	0.24	0.38	0.40	0.58	0.16	0.11
Non–special needs only	0.85	0.79	0.76	–0.09	–0.03	–0.10
Some special needs	0.15	0.15	0.17	0.02	0.01	0.07
All special needs	0.00	0.05	0.07	0.07	0.01	0.15
Grade level	6.12	4.62	4.48	0.62	0.05	–0.19
School average test score	–0.04	–0.19	–0.36	0.46	0.34	–0.45
Percentage of school that is Black/Hispanic	0.37	0.41	0.53	0.45	0.29	0.54

NOTES: This table depicts the covariate balance in one of the five multiple imputation runs of the analysis, prior to the regression stage. The Kolmogorov-Smirnov (KS) statistic is a nonparametric measure of similarity between the distribution of two variables. The Adjusted Difference column is the difference between the group average values in standard deviation units, based on Hedges’ G statistic, except that we adapted the formula for the control group to include the weighted standard deviation and the effective sample size after weighting.