AIR®

AMERICAN INSTITUTES FOR RESEARCH®

Impact of TNTP's Teaching Fellows in Urban School Districts

May 2017

Impact of TNTP's Teaching Fellows in Urban School Districts

May 2017

R. Dean Gerdeman
Yinmei Wan
Ayrin Molefe
Johannes M. Bos
Bo Zhu
Sonica Dhillon

American Institutes for Research

The authors appreciate the collaboration of the following content experts: Jim Lindsay and Andrew J. Wayne.

Contents

Tables

Figures

Summary

In 2010, TNTP received an Investing in Innovation (i3) validation grant from the U.S. Department of Education's Office of Innovation and Improvement to implement and evaluate a version of its Teaching Fellows program. This report summarizes our evaluation of the program's implementation and impacts.

The Teaching Fellows program, an alternative to university-based teacher preparation, was designed to provide urban partner districts with qualified new teachers to fill vacancies and to improve persistently low student achievement. The program recruited and selected trainees from a pool of professionals and recent college graduates. These Teaching Fellows participated in a 6- to 8-week preservice training program and a subsequent inservice training program throughout their first year of teaching. To make certification decisions, TNTP evaluated Fellows on multiple measures, including a series of classroom observations, performance ratings by the principal and, where available, student surveys and student achievement. Fellows who failed to meet performance expectations were not recommended for certification by TNTP.

Under the i3 grant, TNTP implemented the program in seven partner urban school districts that serve large proportions of high-needs students: Baltimore City Public Schools, Charlotte-Mecklenburg Schools, Chicago Public Schools, District of Columbia Public Schools, Fort Worth Independent School District, Metropolitan Nashville Public Schools, and New Orleans public and charter schools.

The evaluation focused on three cohorts of Fellows: those entering classrooms in partner districts in fall 2011, 2012, and 2013. Some components of the evaluation also included Fellows from an initial 2010 cohort in one district. Across these cohorts, TNTP recommended 1,195 Fellows for certification who completed the program and demonstrated mastery of instructional skills to TNTP's standards.

Study Design

The evaluation examined fidelity of implementation using quantifiable indicators of delivery of core program components and Fellows' perceptions about the training they received. TNTP program records and surveys, supplemented by interviews with TNTP staff, provided data for examining program implementation.

The evaluation examined program impacts using quasi-experimental methods. The primary research questions focused on student achievement and teacher instructional practice outcomes in the second year of teaching, after Fellows had completed the program.

- Do students taught by Fellows in the second year of teaching demonstrate higher academic achievement compared with similar students taught by comparable teachers from other certification routes?
- Do Fellows in the second year of teaching demonstrate more effective classroom instructional practice than similar teachers from other certification routes?

Additional research questions focused on student and instructional practice outcomes in the first year of teaching, when Fellows were participating in the program; follow-up analysis of instructional practice

outcomes in the third year of teaching; and comparison of retention rates between Fellows and new teachers from other certification routes. We obtained retrospective and de-identified data from district and state data systems, including student scores from state assessments, student and teacher demographics, school characteristics, and classroom observation scores to address the research questions. These data were used to create comparable groups of Fellows and teachers who did not participate in the Teaching Fellows program, and comparable groups of students of Fellows and students of other teachers.

Student Achievement

To measure impacts on student achievement, we used propensity score matching, in which classes and students taught by Fellows in tested grades and subjects were matched with classes and students taught by comparison teachers in the same districts. This process resulted in two groups of students considered equivalent on prior student achievement (the pretest) and the proportions of English language learners (ELLs), minorities, students eligible for free or reduced-price lunch, and students with individualized education programs (IEPs). Students were combined across districts and cohorts.

The sample to address the primary research question on student achievement included 12,795 students taught by 303 Fellows and 10,778 students taught by 693 comparison teachers in the second year of teaching. The sample represented 29% of all Fellows in the second year of teaching who were in district teacher records across the cohort studied, and between 86% to 94% of Fellows in tested grades and subjects within each cohort. Teachers in the sample were distributed among Grades 4–5 (30%), Grades 6–8 (40%), and Grade 9–12 (32%). These teachers primarily taught in mathematics (45%), reading (43%), science (22%), with a smaller proportion (5%) teaching social studies.[1] The majority of Fellows in district records were not linked with students in tested grades and subjects with available prior achievement records, and thus were excluded from this analysis. Excluded teachers typically taught in Grades K–3 (where students are not routinely tested or lack prior-year assessments), middle and high school subjects not tested each year, or special education.

Standardized assessment scores (*z*-scores), derived from assessments in different grades and subjects, served as the outcome measure and measure of prior achievement. We estimated differences in achievement between students taught by Fellows and comparison teachers using a statistical model that accounted for observable student characteristics, prior achievement, classroom-level characteristics, and teacher characteristics.

Teacher Outcomes

To estimate impacts on instructional practice, we used propensity score matching in which Fellows were matched with comparison teachers in the same districts based on observable teacher (including experience, age, gender, minority status, subject, and grade level) and school (including student demographic characteristics and school performance) characteristics. This process resulted in comparable groups of Fellows and comparison teachers, based on the available data.

[1] The distributions count individual teachers more than once, if teachers taught in more than one grade band or in multiple subjects.

The sample to address the primary research question on teacher instructional practice included 452 Fellows and 789 comparison teachers in the second year of teaching. The sample represented 78% of Fellows in the second year of teaching who were in district teacher records across the cohorts studied. Teachers in the sample were distributed among all grades, including pre-kindergarten (7%), Grades K–5 (42%), Grades 6–8 (18%), high school (26%), and other grade configurations (7%). Nearly half of the teachers (48%) taught multiple subjects, with smaller proportions teaching single subjects (language arts – 14%, mathematics – 11%, science – 9%, social studies – 3%), special education (6%), early childhood (4%), or other areas (4%). There were no sufficiently comparable teachers in the same grades and subjects for 22% of the Fellows in district records, particularly at the high school level and in mathematics, science, and special education; these Fellows were excluded from the sample.

Classroom observation scores from the district teacher evaluation system served as the outcome measure; these data had been collected originally by school administrators or other local observers for the purposes of teacher evaluation and feedback. We examined differences in classroom observation scores between Fellows and comparison teachers using regression analysis that accounted for teacher and school characteristics available in the district data.

As a supplemental component of the evaluation, we used longitudinal analysis to compare retention rates of Teaching Fellows with all other new teachers in their districts.

Findings

The Teaching Fellows program was implemented with fidelity during the evaluation period. All district sites met predetermined evaluation benchmarks for implementation of core program components.

The academic performance of the students of Fellows in the second year of teaching was similar, on average, to students of matched comparison teachers, as measured by standardized achievement scores with students pooled across districts and cohorts ($g = 0.017$,[2] $p = .631$). The estimated differences in academic achievement did not differ significantly by district, grade level, subject, or cohort. Findings were similar for student achievement in the first year of teaching, with no significant differences observed between students of Fellows and students of comparison teachers.

For teacher outcomes, Fellows in the second year of teaching demonstrated similar instructional practice on average to that of comparison teachers, as measured by overall classroom observation scores combined across districts ($g = 0.00$, $p = .942$). There were no significant differences within any of the districts. The findings were similar for instructional practice in the first and third years of teaching, though the change in estimated differences between Fellows and comparison teachers across years suggests a relative improvement for Fellows over time.

The retention rate for Fellows in teaching positions into the second year was 6 percentage points higher than for other new teachers in their districts, a statistically significant difference. The Fellow cohorts examined included approximately 8% of Fellows per year who ultimately failed to meet TNTP's

[2] *g* stands for Hedges' *g*, which represents standardized mean group difference.

performance expectations and were not recommended for certification. If these Fellows had been excluded from the retention analysis, the positive difference in retention for Fellows likely would have been larger. Data limitations prevented us from accurately identifying nonpassing Fellows in the analysis.

Conclusions

The study found that students taught by Fellows performed similarly to students taught by comparable teachers and that Fellows demonstrated similar classroom instructional practice to comparable teachers. In the districts and Teaching Fellows cohorts examined, TNTP recommended certification for approximately 1,200 new teachers, who were retained into the second year of teaching at higher rates than other new teachers. These Fellows were found to have similar performance even though their training period was shorter than is typically provided through traditional teacher preparation programs. Considering the persistent need for qualified teachers in urban districts, the study indicates that the Teaching Fellows program recruited and trained qualified teachers and provided a viable pathway for new teachers in the partner districts. However, the findings suggest that TNTP fell short of its goal to produce a cadre of new teachers who were more effective than other new teachers hired to fill vacancies in the study districts.

Relative to prior research on TNTP's Teaching Fellows programs, the methods used in this study offered several advantages, including use of data from all districts involved in a large-scale implementation of the program; statistical methods to create comparable groups of teachers with the same level of experience; and outcome measures focused on teacher practice as well as student achievement. The methods used also had several important limitations. First, it is possible that the groups differed on variables that were not available, such as teacher attributes that TNTP might have considered in recruitment; the findings cannot be isolated from any unmeasured differences. Second, only the subset of Fellows in tested grades and subjects were included in the analysis of student achievement, which may limit generalizability to Fellows who teach in other areas. And third, the teacher observation measures were originally collected for a different purpose and had variable degrees of reliability and consistency, potentially affecting the likelihood of observing differences in this study.

TNTP's Teaching Fellows Program

Alternative Certification Programs Were Developed in Response to Teacher Pipeline Needs

The most influential school-based factor affecting student achievement is teacher quality (Aaronson, Barrow, & Sanders, 2003; Chetty, Friedman, & Rockoff, 2011; Goldhaber, 2002; Gordon, Kane, & Staiger, 2006; Hanushek, Kain, & Rivkin, 1998; Nye, Konstantopoulos, & Hedges, 2004; Sanders & Horn, 1998). Students who have effective teachers have significantly greater learning gains than those with ineffective teachers (Nye et al., 2004). In fact, having effective teachers may be particularly important for students who are the most disadvantaged (Gordon et al., 2006). Yet in some districts, schools serving low-income and minority students, which struggle with low student achievement and high dropout rates, are less likely to be staffed with effective teachers (Clotfelter, Ladd, & Vigdor, 2007; Coopersmith, 2009; DeAngelis, Presley, & White, 2005; Isenberg et al., 2016; Lankford, Loeb, & Wyckoff, 2002; Sanders & Rivers, 1996). Improving teacher quality for all students, regardless of background, is critical to improving student academic growth and closing achievement gaps.

Education researchers and policymakers also have been concerned about shortages of elementary and secondary teachers in American schools (Ingersoll, 2003; Ingersoll & May, 2012; Miller & Chait, 2008). Bilingual education and English language acquisition, foreign language, mathematics, science, and special education have been consistently designated as statewide teacher shortage areas in many states (U.S. Department of Education, 2015). The reasons for teacher shortages are many and complex, but students, especially those with learning gaps, suffer from disruptions caused by persistent teacher shortages (Jacobs, 2007).

Alternative routes to teacher certification emerged in the early 1980s as a way to address existing or projected shortages of teachers and their potentially harmful effect on student learning. These routes—"alternatives" to traditional teacher education programs—allow prospective teachers to begin teaching more quickly while completing training and mentoring requirements. The number of teachers obtaining certification through alternative routes has increased substantially in the past 25 years, from several hundred in the mid-1980s to about 60,000 in the late 2000s (Feistritzer, 2011). It is estimated that alternative certification programs supply one of every five teachers in the United States (Greenberg, McKee, & Walsh, 2013). Large urban districts in particular—such as Chicago, New Orleans, and New York City—increasingly rely on alternative preparation programs to provide a significant number of new teachers (Barrett & Harris, 2015; Grossman & Loeb, 2010; Ng & Peter, 2010).

Research indicates that alternative certification programs can increase the supply of teachers and change the composition of the teacher workforce (McKibbon, 1998; Shen, 1999; U.S. Department of Education, 2016; Wilson, Floden, & Ferrini-Mundy, 2001). Research comparing the effectiveness of alternatively certified teachers with that of traditionally trained teachers has yielded mixed results. Some studies have found that alternatively certified teachers are as effective as traditionally trained teachers (Clark et al., 2015; Constantine et al., 2009; Harris & Sass, 2007), and others have found effects favoring alternative certification programs (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2006;

Clark et al., 2013). Given the importance of alternative routes in the teacher workforce, further evidence is needed about whether and how specific alternative certification programs are supporting improved outcomes, especially for underserved students in urban communities.

TNTP's Programs Provide Alternative Pathways to the Teaching Profession

Founded in 1997 as The New Teacher Project, TNTP aims to attract capable individuals into teaching and provide candidates with sufficient preservice and on-the-job training and coaching to improve student outcomes. Early in its history (1997–2002), TNTP partnered with urban districts, including Baltimore, the District of Columbia, New Orleans, and New York City, to help these districts improve the way they recruited, trained, and hired new teachers.

In 2000, TNTP began the Teaching Fellows program and the Practitioner Teacher program. The Teaching Fellows program is an alternate route to certification that employs a rigorous selection and training process to provide teachers for shortage subject areas and high-poverty schools. The Practitioner Teacher program is an independent, state-licensed certification program that is specifically designed to prepare teachers to raise student achievement in high-poverty schools. As TNTP expanded its programs to more locations and became increasingly familiar with the needs of urban districts, it has refined the programs to improve teachers' effectiveness and help partner districts address their unique challenges.

TNTP has become one of the largest suppliers of teachers in the country. As of 2015, TNTP had partnered with districts and charter management organizations in more than 30 cities and had trained approximately 34,000 teachers for high-need schools and hard-to-staff subjects. In some urban districts such as New York City, TNTP teachers have accounted for one fourth or more of new teachers hired within a year. TNTP is also a state-approved certification provider in seven states and the District of Columbia.

More Evidence on the Impact of TNTP's Program Is Needed

Three studies analyzed the relative effectiveness of Teaching Fellows using panel data from single TNTP sites. Using 6 years (1998–99 through 2003–04) of data on test performance for students in Grades 3–8 in the New York City public schools, Boyd et al. (2006) examined the effectiveness of teachers from different certification pathways into teaching in New York City, including comparing students of Fellows to students of traditionally prepared teachers with the same level of experience. The study found that students of first-year Fellows at the elementary level had slightly lower scores in math and reading (by 0.04 *SD* [standard deviation] in both subjects), while students of first-year Fellows at the secondary level had similar scores in mathematics and slightly lower scores (by 0.03 *SD*) in reading. By the third year of teaching, there were no observed differences in scores between students of Fellows and traditionally prepared teachers at the elementary level, and students of Fellows had higher scores in both subjects at the secondary level.

In a different study using 1998–99 through 2004–05 data from the New York City public schools, Kane et al. (2006) found that students taught by Fellows performed as well as students taught by

traditionally prepared teachers in mathematics and slightly lower in reading. The study also indicated that Fellows made gains over the first 3 years of teaching, with relative differences between Fellows and traditionally prepared teachers becoming slightly more positive in both subject areas. All differences observed in this study were small, with effect sizes smaller than 0.02 *SD*.

Noell et al. (2009) assessed the impact of recent graduates from specific teacher preparation programs on academic achievement of Louisiana students in Grades 4–9 using a pooled data set spanning the academic years 2004–05 to 2007–08. The study found that students taught by TNTP teachers outperformed students taught by experienced teachers trained through other routes on the state mathematics assessments (by 0.11 *SD*) and reading assessments (by 0.08 *SD*).

To account for differences in the students taught and in the work environments in which they taught, all three studies (Boyd et al., 2006; Kane et al., 2006; Noell et al., 2009) employed quasi-experimental designs that statistically controlled for a number of student-, classroom-, grade-, and school-related factors and teachers' experience levels; and used multiple model specifications to check for the robustness of the findings. However, there is evidence in those studies that teachers from different preparation pathways in the samples also differed in the characteristics of their students. For example, Fellows in general were found to be more likely to work with nonwhite, low-income, and low-performing students than traditionally prepared teachers. Statistical controls alone may not adequately account for the imbalances between the Fellow and comparison groups in quasi-experimental designs.

A more recent study by Clark et al. (2013) provided stronger evidence on the Teaching Fellows program using a rigorous experimental design.[3] Within 44 participating schools across eight districts, students in Grades 6–12 were randomly assigned to mathematics classrooms taught by either a Fellow or a comparison teacher who did not participate in a highly selective alternative certification program. The study found that, on average, students of Fellows and students of comparison teachers had similar scores on end-of-year state mathematics assessments. Clark et al. (2013) used random assignment to ensure that there were no systematic differences at the start of the school year between students assigned to Fellows and those assigned to comparison teachers. Fellows in the study were considerably less experienced in teaching (with an average of 4 years) than other secondary mathematics teachers in their schools (with an average of 13 years), though supplemental analyses of the subgroups of teachers with equivalent amounts of teaching experience also indicated no differences.

These prior studies provide useful evidence about the impact of TNTP's Teaching Fellows program but with several limitations. The research has focused on student achievement outcomes; none of the studies highlighted above systematically examined teacher instruction or teacher outcomes. Comparison groups used in prior studies differed in some observable characteristics, which suggests that the study designs may not have adequately accounted for selection biases or potentially important baseline differences between teachers or students. The prior studies also were conducted in contexts that may limit generalizability, considering that TNTP operates Fellows programs in many communities with different local conditions and partners. The three quasi-experimental studies were limited to two

[3] This study met What Works Clearinghouse (WWC) evidence standards without reservations, an indication of a high-quality design for a randomized controlled trial.

sites, New York City and greater New Orleans. The Clark et al. (2013) study was conducted in districts and schools that were willing to participate in a randomized controlled trial; these schools were found to differ on some dimensions (e.g., charter status and urbanicity) from all secondary schools with Teaching Fellows placements nationwide. Additional rigorous research is needed to provide more evidence about the effectiveness of TNTP programs on a broader set of outcomes and in varied settings, as well as to document TNTP's approach to recruiting, preparing, and certifying teachers.

The Teaching Fellows Program Was Designed to Provide Effective New Teachers for Urban Schools

TNTP was awarded a selective Investing in Innovation (i3) validation grant from the U.S. Department of Education in 2010 to implement and validate a version of the Teaching Fellows program. American Institutes for Research (AIR) conducted an independent evaluation of the program.

Under the i3 grant, the Teaching Fellows program was a multisite teacher pipeline initiative that paired TNTP's Teaching Fellows program and the Practitioner Teacher program to create a more integrated and potentially impactful alternative pathway for new teachers.

TNTP standardized its recruitment, selection, training, and certification services for all sites. The program encompassed the following core components, displayed in Figure 1:

- Recruitment and selection of Fellows, which involved identifying promising teacher candidates through outreach to experienced recent college graduates and career changers who lack prior teaching experience
- Preservice training for Fellows, which involved participating in a 6- to 8-week summer institute and summer teaching experience and a screening at the end of training to identify those who showed the potential to meet the expectations of TNTP for effective teaching
- Inservice training for Fellows, which involved participating in seminars and coaching during the first year of teaching (previously called the Practitioner Teacher program)
- The Assessment of Classroom Effectiveness (ACE) that used multiple measures to identify potentially effective teachers to recommend for certification at the end of the inservice training

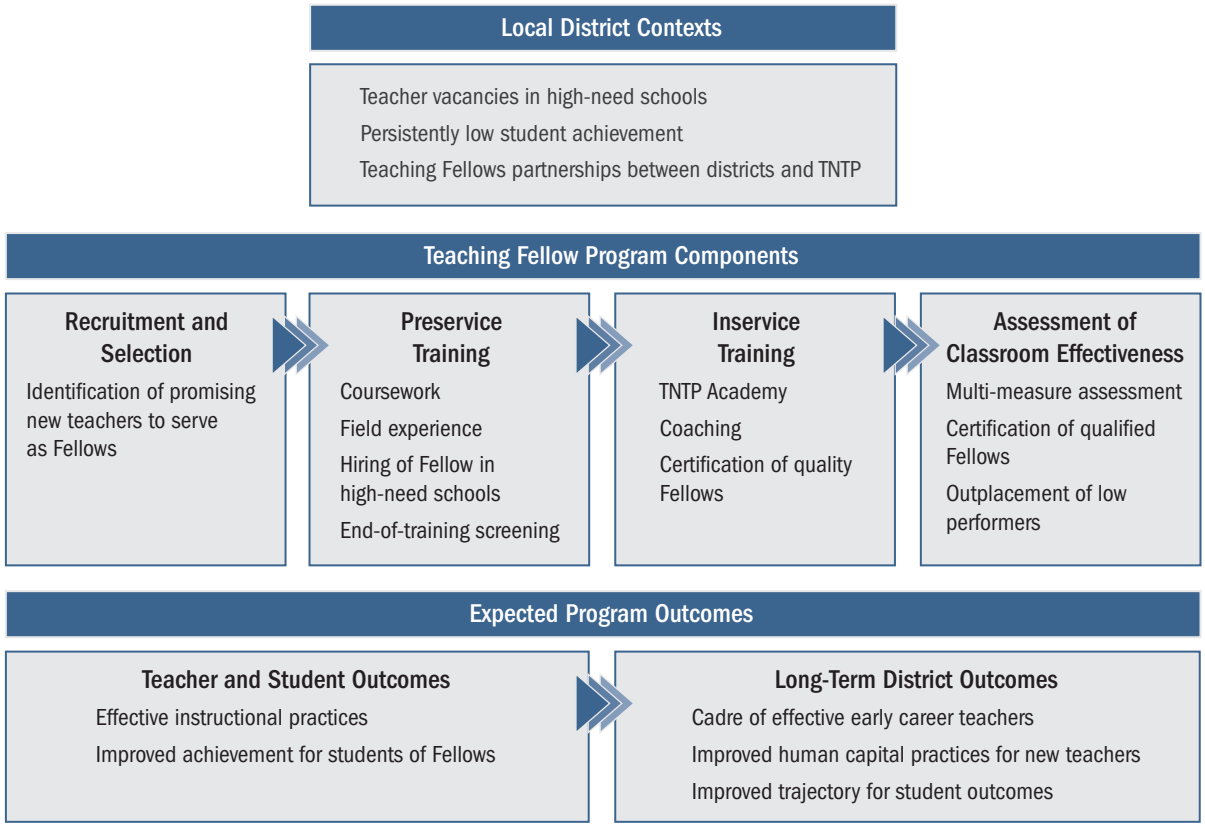
A key innovation in the TNTP program was the incorporation of effectiveness screening at two points in the program. Participants were required to pass a screening at the end of their summer preservice training and to pass another screening (the ACE) at the end of their first year of teaching.

The program was designed for delivery through a distributed organizational structure, with site-based teams of two to four TNTP staff (including a site manager) who worked with human capital staff in their partnering district. Although all sites used TNTP's overall program model, each site had its own semiautonomous implementation team, which was intended to enable TNTP staff to identify recruitment needs, set recruitment targets, and make ongoing adjustments to accommodate the specific needs and circumstances of each district. Costs to Fellows for participating in the program varied across districts, with Fellows typically paying about $4,000 to $6,000 in fees. (See Appendix A for more information on the program.)

Through the Teaching Fellows program, TNTP provided partnering districts with annual cohorts of Fellows to fill vacancies in schools across the district. However, the program was not designed only to provide a pipeline of new teachers. The combination of selecting high-quality candidates and subsequent training, including demonstration of competency on performance assessments, was designed to produce program completers who were more effective in improving student achievement than typical new teachers from other types of preparation programs who might fill teaching vacancies. TNTP developed the program components with an expectation that Fellows who were less effective or less committed would depart as they progressed through the program.

As illustrated in Figure 1, TNTP expected that program completers would demonstrate improved outcomes compared with those of other new teachers, including instructional practice and student achievement. Over time, TNTP expected that the program would provide a cadre of effective early career teachers and support local districts' efforts to improve their practices for training new teachers.

Figure 1. Overview of the Teaching Fellows Program

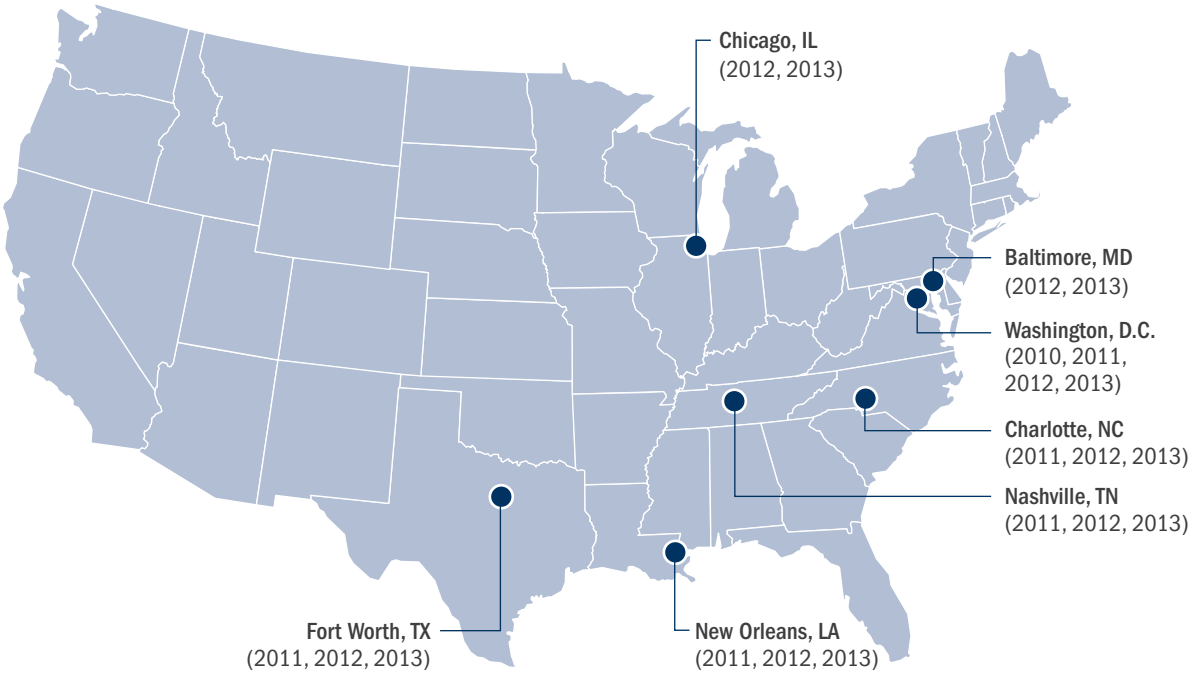


TNTP Partnered With Seven Urban Districts to Implement the Program

Under the i3 grant, the Teaching Fellows program was implemented in seven large urban school districts.[4] Each district met at least two of three criteria established by TNTP for identifying high-need students: serving 60% or more students from minority racial/ethnic groups; serving 50% or more students eligible for free or reduced-price lunch; or having a higher percentage of students not meeting adequate yearly progress goals than the state average.

The grant funded program implementation for one small pilot cohort of Fellows who started in 2010 and larger cohorts of Fellows who started in 2011, 2012, and 2013. As shown in Figure 2, Washington, D.C., launched a 2010 cohort, five sites launched a 2011 cohort, and all seven sites launched 2012 and 2013 cohorts.

Figure 2. Teaching Fellows Cohorts and District Sites Included in the Evaluation



Note. The parentheses in the figure represent the cohorts of Teaching Fellows that were examined in this evaluation.

[4] The districts implementing the Teaching Fellows program funded by the i3 grant included Baltimore City Public Schools, Chicago Public Schools, Charlotte-Mecklenburg Schools, District of Columbia Public Schools, Fort Worth Independent School District, Metropolitan Nashville Public Schools, and New Orleans public and charter schools. In this report, these sites are referred to by the city names.

The cohorts represent the years in which Fellows entered the program, participated in preservice training in the summer, and typically entered inservice training as teachers employed by the districts in the fall. TNTP had previously operated its Teaching Fellows program or Practitioner Teacher program in five of these districts. With the i3 grant, TNTP created or expanded the combined Teaching Fellows program and Practitioner Teacher program in these five sites and at two new sites.[5]

To select participating districts, TNTP identified districts that met the eligibility criteria for the i3 grant and were interested in using teacher recruitment and teacher evaluation as a way to potentially address these gaps. TNTP recruited six of the seven districts through bilateral discussions with each district's leadership prior to submitting the i3 grant proposal. After the grant was awarded, TNTP added Baltimore as a seventh district.

Only Fellows Who Met Multiple TNTP Milestones Completed the Program

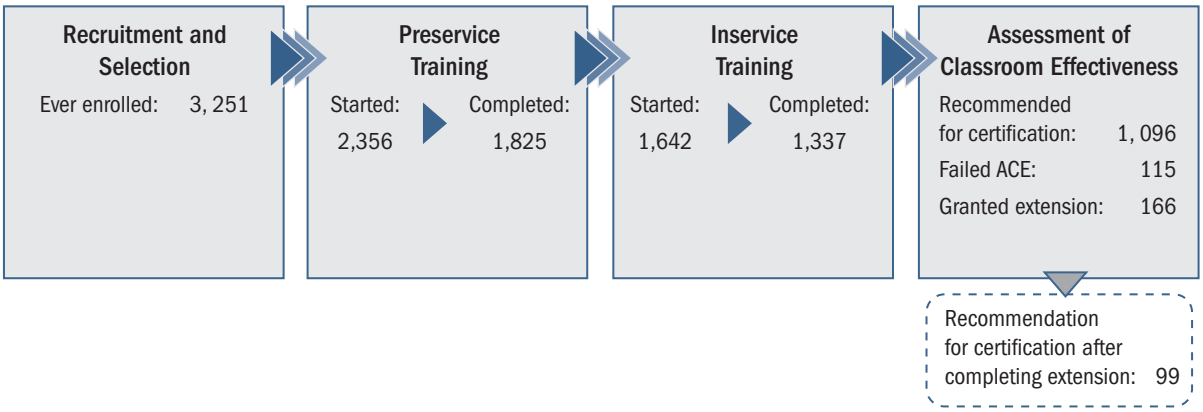
As summarized previously, Fellows were required to participate in several steps toward their eventual certification and placement in a teaching job. Some Fellows dropped out at each step, which is an expected part of the program design (TNTP, 2014). Figure 3 shows the number of Fellows who reached major program milestones. The distributions by site are presented in Table A1.

Across cohorts and district sites examined in the evaluation, TNTP enrolled 3,251 Fellows. Of the 3,251 enrollees, 1,825 (56%) completed the preservice training components. The noncompleters consisted of individuals who agreed to participate but did not actually start training, those who participated but failed to meet expectations of the end-of-training screening and were counseled out by TNTP staff, and those who decided to leave the program on their own after starting training. Of the 1,825 Teaching Fellows who completed preservice training, 1,642 (90%) successfully obtained teaching positions in the partner districts and entered TNTP's inservice training.

The majority of Fellows who entered inservice training (1,377 or 84%) completed the inservice training and a full year in the classroom. As expected by TNTP, a small proportion of Fellows (115 individuals, representing 8% of completers) failed to demonstrate adequate progress on the ACE and thus were not recommended for certification by TNTP. A larger group (166 individuals, representing 12% of completers) were given an extension year. The rest of the inservice training completers (1,096 individuals representing 80% of completers) passed the assessment and were recommended for certification from TNTP. Among the 166 who were granted an extension year, 99 (or 60%) ultimately were recommended for certification in the second year after completing the extension plan and successfully passed the assessment. In sum, 1,195 Fellows were recommended for certification.

[5] A TNTP-district partnership dated back to 2001 for District of Columbia Public Schools, 2002 for Baltimore City Public Schools, 2006 for Chicago Public Schools, 2007 for New Orleans Recovery School District, and 2009 for Metropolitan Nashville Public Schools. TNTP created first-time partnerships with Charlotte-Mecklenburg Schools and Fort Worth Independent School District with the i3 grant.

Figure 3. Number of Fellows Who Reached Each Program Milestone



Source: American Institutes for Research's analysis based on program records provided by TNTP. The figure presents data for the Fellows in the 2011, 2012, and 2013 cohorts across districts. Comparable data were not provided for the 2010 pilot cohort.

TNTP Implemented the Program With Fidelity in All Districts

We examined the fidelity of program implementation using common and quantifiable indicators across the cohorts and district sites, as summarized in Appendix A. The measurement of implementation fidelity involved two categories of indicators: (a) site-level indicators of the extent to which core components of the program were delivered as originally intended by TNTP (e.g., the candidate selection process, Fellows' hours of participation in program activities, and Fellow assessment processes) and (b) Fellows' responses to surveys on their perceptions about the training they received from coaches and trainers. The resulting measure of implementation fidelity consisted of concrete and measurable indicators for each of the four Teaching Fellows program components: recruitment and selection, preservice training, inservice training, and the Assessment of Classroom Effectiveness.[6]

Each indicator had a designated threshold for adequate implementation. The data for each indicator were analyzed to determine whether a site met the threshold, then scores across indicators were aggregated to measure whether each program component met implementation thresholds established for the evaluation, within and across sites and cohorts (see Appendix A for further explanation).

All district sites met the implementation threshold for each program component, as established by this evaluation. These findings indicate that TNTP was successful in implementing the major components of the program across various districts and contexts, and over multiple years, consistent with expectations under the i3 grant.

[6] The implementation indicators were adjusted to respond to changes introduced by TNTP for the 2012 and 2013 cohorts, including a revised preservice curriculum and integration of coaching (see Appendix A for more details).

Methods

Questions and Data Sources

The Study Examined Impacts on Student Achievement and Teacher Instructional Practice Using Matched Comparison Groups

The evaluation focused on estimating impacts of the Teaching Fellows program in the second year of teaching, after Fellows had completed the program:

- Do students taught by Fellows in the second year of teaching demonstrate higher academic achievement compared with similar students taught by comparable teachers from other certification routes?
- Do Fellows in the second year of teaching demonstrate more effective classroom instructional practice than similar teachers from other certification routes?

Supplemental research questions focused on teachers' first year of teaching, when Fellows were participating in the program, and on teachers' third year of teaching as a follow-up component:

- Do students taught by Fellows in the first year of teaching demonstrate higher academic achievement compared with similar students taught by comparable teachers from other certification routes?
- Do Fellows in the first year of teaching demonstrate more effective classroom instructional practice than similar teachers from other certification routes?
- Do Fellows in the third year of teaching demonstrate more effective classroom instructional practice than similar teachers from other certification routes?

These questions were addressed using quasi-experimental methods that relied on matching to create comparison groups that were comparable to Fellows and their students

The Study Compared Teacher Retention Rates Between Fellows and Other New Teachers Using Longitudinal Analysis

A separate component of the study focused on teacher retention:

- Are Fellows retained in teaching positions in their districts at a greater rate than other new teachers in the second, third, and fourth year of teaching?

This question was addressed using longitudinal analysis to compare retention in teaching positions, within districts, between Fellows and other new teachers who started in the same years.

Existing District Data Were Used to Address the Questions

We obtained existing data from participating districts to evaluate program impacts. For the analysis of student achievement, we obtained data on students and teachers in tested subjects and tested

grades with relevant prior-year test scores. These data included students' test scores in state assessments, student demographics, teacher demographics, grades and subjects taught, class rosters, and school-level proficiency rates. For the analysis of teacher instructional practices and teacher retention, we obtained classroom observation scores, teacher demographics, grades and subjects taught, and school-level student demographic characteristics for teachers in all subjects and grades within the districts. (See Appendix B for more information on the instructional practice measures.)

Data for academic years 2010–11 through 2014–15 were included in the evaluation, with different years of data included for different district sites, depending on when TNTP implemented Fellow cohorts in each site and what data were provided by districts (Table 1). For the research questions on student achievement, we obtained necessary data for all participating sites. For the research questions on instructional practice, we obtained necessary data from four districts with eligible teacher observation rubrics. (See Appendix C for more information.)

Table 1. Teaching Fellows Sites and Cohorts Included in the Evaluation, by Outcome

District	2010 Fellows Cohort		2011 Fellows Cohort		2012 Fellows Cohort		2013 Fellows Cohort	
	Student achievement	Teacher instruction	Student achievement	Teacher instruction and retention	Student achievement	Teacher instruction and retention	Student achievement	Teacher instruction and retention
Baltimore					✓	X	X	X
Charlotte			✓	X	✓	X	✓	X
Chicago					✓	✓	✓	✓
District of Columbia	✓	✓	✓	✓	✓	✓	✓	✓
Fort Worth			✓	X	✓	X	✓	X
Nashville			✓	✓	✓	✓	✓	✓
New Orleans			✓	✓	✓	✓	✓	✓[a]

Note. Data from different combinations of sites and cohorts were included to address different research questions, based on available data.

[a] The 2013 cohort in New Orleans was not included in the retention analysis due to data limitations.

✓ = Site implemented Teaching Fellows program and provided necessary data for this analysis.

X = Site implemented Teaching Fellows program but either did not have the necessary data available or did not provide the necessary data for this analysis.

Grey shading = Site did not implement Teaching Fellows program for this cohort and school year.

Box 1. Characteristics of All Fellows Versus All New Teachers in Their Districts

TNTP's Fellows constituted, on average, 6% to 8% of all new teachers across districts for the 2011, 2012, and 2013 cohorts. Among the districts examined, Fellows had the highest representation among new teachers in the District of Columbia (16%) and lowest representation in Chicago (3%). Compared to the broader population of new teachers in their districts, Fellows were younger by an average of approximately 2 to 5 years, and somewhat more likely to be male (30% versus 26%), and White (67% versus 63%).

Methods for Estimating Impacts on Student Achievement

Student Achievement Was Measured Using Standardized Scores From State Assessments

State annual standards-aligned assessments in tested subjects (mathematics, reading, science, or social studies) and tested grades (Grades 3–12) were used as measures of student achievement, including both prior achievement and achievement in the academic year being examined. To facilitate pooling of samples across districts, cohorts, grades, and subjects, we converted students' test scores into standardized scores (z-scores). Standardized scores were calculated as the difference between a student's raw score and the district average raw score for a particular assessment (in a given subject, grade, and cohort), divided by the district standard deviation of raw scores for that assessment. The pooling of samples across subjects, grades, and cohorts was done to maximize statistical power to detect impacts.

Students of Fellows in Tested Grades and Subjects Were Matched With Students of Teachers Trained Through Other Programs

In this evaluation, the quality of evidence about the impact of the Teaching Fellows program on achievement was dependent on identifying a sample of comparison teachers and students who closely resembled Fellows and their students in background characteristics. Drawing on the data provided by the study districts, AIR used background characteristics that are known correlates of student achievement and program participation as a basis for choosing matched comparison groups.

As a first step, we imposed sample inclusion restrictions by limiting comparison teachers to only those who were trained through other programs[7] and who had the same years of experience as Fellows, according to district records (i.e., 0 years of prior experience for teachers in their first year of teaching, 1 year of prior experience for teachers in their second year of teaching). Research shows that teachers' impact on student achievement improves during the first several years of teaching (Harris & Sass, 2007; Rockoff, 2004), making it important to balance teaching experience between the Fellows and comparison groups.

[7] TNTP and Teach For America (TFA) were typically the largest providers of alternative certification in the urban districts. The pool of potential comparison teachers included teachers in district records who were not identified as TNTP or TFA teachers. TFA corps members participated in TNTP inservice training at some sites and thus were excluded from the comparison pool as much as possible based on available data. The districts did not provide data on how teachers in the comparison pool were prepared or certified.

Next, we used propensity score matching to identify comparison teachers based on teacher demographic and classroom characteristics and to identify comparison students based on prior student achievement and other student characteristics.[8] The matching process was implemented in three steps and was conducted separately for each combination of cohort, district, grade, and subject. First, for every cohort within a district, each class taught by a Fellow was matched with up to two classes in the same grade and subject area taught by comparison teachers with the closest propensity scores (i.e., the classes with the closest propensity to being taught by a Fellow). Next, students were matched within each group of matched classes; that is, each student taught by a Fellow was matched with up to two students taught by a comparison teacher with the closest propensity scores (i.e., the students with the closest propensity to being in a Fellow's classroom). After matching, the matched samples were combined across cohorts, districts, subjects, and grades.

As a final step, we checked whether matching produced two groups that were equivalent in observed background characteristics. Consistent with What Works Clearinghouse (WWC) standards (WWC, 2013), we considered the two groups to be balanced if the standardized average difference in prior student achievement (the pretest measure) between the two groups of students in the combined sample was less than or equal to 0.25 *SD*. (See Appendix D for more details.)

The matching process produced two analytic samples (see Appendix E for details):

- The sample to address the primary research question (impacts in the second year) consisted of 23,573 students taught by teachers in their second year of teaching: 12,795 taught by 303 Fellows and 10,778 taught by 693 comparison teachers.[9] Fellows in this sample collectively represented 29% of the Fellows from the 2010, 2011, 2012, and 2013 cohorts who were in districts' teacher records in their second year of teaching (see Figure 4).[10]
- The sample to address the supplemental research question (impacts in the first year) consisted of 37,099 students taught by teachers in their first year of teaching: 18,826 taught by 445 Fellows and 18,273 taught by 1,014 comparison teachers. Fellows in this sample collectively represented 37% of the Fellows from the 2011, 2012, and 2013 cohorts who were in districts' teacher records in their first year of teaching (see Figure 4).[11]

---

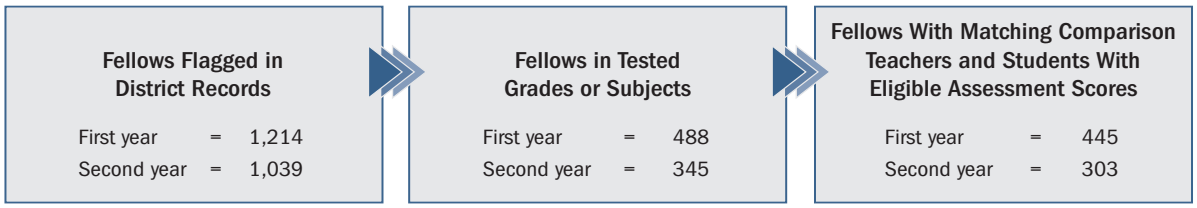
[8] Variables used for matching classes included teacher's age, racial minority status and gender, and the following classroom characteristics: average of students' prior-year test scores (z-scores), grade level, subject, percentage of ELL students, percentage of minority students, percentage of students eligible for free or reduced-price lunch, percentage of students with an IEP, and percentage of female students. Variables used for matching students included age, ELL status, IEP status, free or reduced-price lunch eligibility, gender, and racial minority status. Several variables (teacher's age and minority status, and student's age and free or reduced-price lunch eligibility) were not provided by all districts. Variables were used for matching within a district when the data were available for that district; variables were used in the impact models only if they were available from all districts.

[9] We implemented the matching of students with replacement—that is, students of Fellows were matched to at most two comparison students with the nearest propensity scores even if one or both of those comparison students had already been previously matched. This method allowed for student matches with more similar propensity scores and resulted in analytic samples with a larger number of students of Fellows than comparison students.

[10] Of the 1,039 Fellows from the 2010, 2011, 2012, and 2013 cohorts who were in the district files in their second year of teaching, 345 Fellows (33%) remained after those who taught nontested grades and subjects were removed. Of these remaining Fellows, 303 (88%) were matched. Among the Fellows who were in records in both the first and second years of teaching, there was a net outflow from tested grades and subjects, with a net of about 3% of Fellows moving from tested to nontested teaching assignments.

[11] Of the 1,214 Fellows from the 2010, 2011, and 2012 cohorts who were in the district files in their first year of teaching, 491 Fellows (40%) remained after those who taught nontested grades and subjects were removed. Of these remaining Fellows, 445 (91%) were matched.

Figure 4. Sample of Fellows for Analysis of Student Achievement Outcomes



Box 2. Fellows' Movement in the Analytic Samples Between the First and Second Years

The samples of second-year teachers used for analyzing student achievement included individual teachers who were not part of the first-year samples. Of the 2011, 2012, and 2013 cohorts of Fellows who were in the selection pool of first-year teachers in tested grade and subjects, approximately 53% also were in the selection pool of second-year teachers in tested grades and subjects, and 26% were identified in the district data files in the second year but were not teaching in tested grades and subjects. About 21% of Fellows who had been in the first-year selection pool were not identified in district data files in the second year, indicating they were not employed as teachers in these districts in that year. (The data obtained for the initial 2010 cohort—in the one participating district that year—did not permit calculating percentage of stayers and in-movers between the first and second years of teaching.)

Analyses Estimated Differences in Student Achievement

For each of the two analytic samples, differences in achievement were estimated using a three-level statistical model (students nested within classes within teachers) that accounted for correlations among students who attended the same classes and were taught by the same teacher, as well as students' demographic characteristics and prior achievement, classroom-level characteristics, and teacher demographic characteristics (see Appendix D for more details). This method produced district-specific estimates of the difference in average achievement between students of Fellows and comparison students.[12] For each sample, the site-specific estimates were then pooled into a weighted average, using the proportion of Fellows in each district as the weight, to generate an estimate of the overall difference in student achievement.

[12] As is customary, this method controlled for variation in available teacher background characteristics in the process of matching Fellows to comparison teachers and in the process of estimating impacts. The resulting impact estimates may be biased if TNTP recruited inherently more qualified teachers and if those qualifications were correlated with the teacher background characteristics controlled for in the study. However, the demographic similarity between Fellows and other new teachers shown in Box 1 makes this type of bias unlikely.

Methods for Estimating Impacts on Teacher Instructional Practice

Teacher Instructional Practice Was Measured Using Classroom Observation Scores From Local Teacher Evaluation Systems

The classroom observation components of the state or district teacher evaluation systems served as the outcome measures for teacher instructional practice. Teacher scores on these classroom observation rubrics were originally collected by principals or other observers as part of their districts' teacher evaluation systems. These classroom observations are used by the district or state to rate or categorize the performance of teachers and provide feedback to teachers.

Observation scores from teacher evaluation systems provide face validity, with directly observable practices reflecting standards that have been adopted by the states or districts (Cohen & Goldhaber, 2016). However, classroom observation scores have been criticized for not adequately differentiating among teachers (Anderson, 2013; Cohen & Goldhaber, 2016). The Measures of Effective Teaching project (Kane & Staiger, 2012) investigated five observation instruments—including the commonly used Framework for Teaching (Danielson, 1996)—and found that scores from these instruments were positively associated with student achievement gains and that averaging scores over multiple observations would likely better assess instructional quality. However, this research was not conducted within the context of a potentially high-stakes teacher evaluation system. Another study examined data from teacher evaluation systems in four mid-sized urban school districts and found that teachers' evaluation scores—a combination of classroom observations, student achievement, and other student and administrator ratings—meaningfully assess teacher performance within a range of reliability and validity consistent with data collected for research studies (Whitehurst, Chingos, & Lindquist, 2014).

Since there is limited published research on the measurement properties of observation scores from local teacher evaluation systems, we considered both the local practices for conducting the observations and the properties of the actual data in determining suitability of the scores as an outcome measure in this evaluation.

All seven district sites included classroom observations in their teacher evaluation systems. We used scores from only the districts that reported the following practices:

- Observations were performed in a systematic way, using protocols and procedures that were standardized or consistent across schools.
- Observers were trained to perform observations.
- Multiple observations were performed per year for individual teachers.
- Observation scores were based on a scale that had gradations of performance beyond just "satisfactory or unsatisfactory."

Five participating sites (Baltimore, Chicago, the District of Columbia, Nashville, and New Orleans) had observation protocols and processes consistent with these criteria. Four of these sites, excluding Baltimore, provided de-identified data for the evaluation. The protocols in place in each site were derived or adapted from Charlotte Danielson's Framework for Teaching (1996). (Appendix B provides more detail on the measures in place in each site.)

We conducted initial analysis of the observation data obtained from each participating district. Teachers' scores demonstrated meaningful variability across teachers, with standard deviations comparable to observations collected for research purposes.[13] Teachers' scores on the indicators showed moderate to high correlations across standards, indicating internal consistency among the standards in each district's rubric, and moderate within-year stability across observations (see Table B1). This analysis thus indicated that the observation scores from the four districts were adequate for use as outcome measures in this evaluation.

The overall observation scores (based on scores from all domains) were used as the primary measure of instructional practice. In addition, all four sites had classroom observation components related to the Instruction domain and the Classroom Environment domain on the Framework for Teaching. Because the Teaching Fellows program focused on skills related to these two domains, AIR used the average observation scores for each domain, when available, as secondary measures of instructional practices.[14]

Box 3. Instructional Practice Scores for New Teachers Versus All Teachers

Teachers in their first year of teaching had lower instructional practice scores compared with all teachers in their district. The average standardized scores for all first-year teachers in the district data ranged between −0.54 and −0.25 *SD*, indicating lower scores than the district means (set to 0) for all teachers. For second-year teachers, average instructional practice scores were much closer to the district mean, between −0.17 and −0.05 *SD*. The findings are consistent with the consensus in the literature that experienced teachers are in general more effective and that new teachers improved effectiveness as they gained experience (Boyd et al., 2006; Clotfelter, Ladd, & Vigdor, 2007; Kane, Rockoff, & Staiger, 2006).

Fellows Were Matched With Teachers Trained Through Other Types of Programs

The data obtained from districts were used to identify Fellows and a sample of comparison teachers from other training programs with similar characteristics to Fellows. As a first step, similar to the approach used for analysis of student achievement, we restricted eligibility to include only the comparison teachers with the same years of experience as Fellows.

The next step in identifying comparable samples of teachers involved identifying groups of comparison teachers who closely resemble Fellows in observable teacher demographic characteristics, teacher grade and subject, and school demographic characteristics.[15] We used a propensity score matching process to identify comparison teachers for each cohort of Fellows within each school district.

---

[13] For example, the standard deviations of observation scores using the Framework for Teaching from the MET study ranged from 0.30 to 0.42 on a 4-point scale (Garrett & Steinberg, 2015).

[14] The findings on domain scores were based on analyses of data in Chicago, the District of Columbia, and Nashville only. Observation data for New Orleans were not disaggregated into domains.

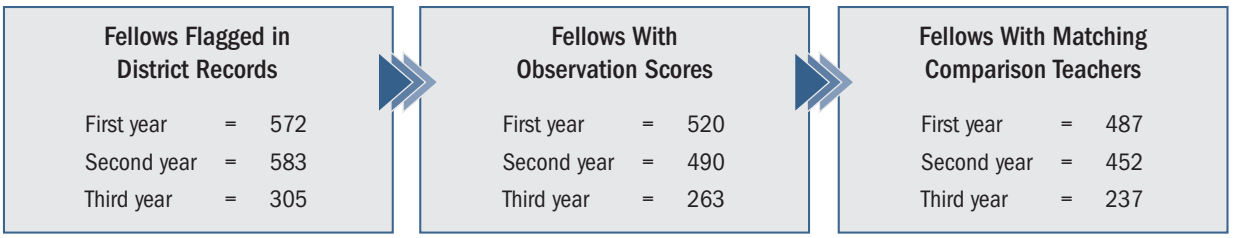
[15] Teacher characteristics included age, gender, minority status, grade level, and subject area. School characteristics included school size, the percentage of students eligible for free or reduced-price lunch, the percentage of ELL students, the percentage of special education students, the percentage of minority students, and the percentage of students proficient in reading and in mathematics on the most recent state standardized tests.

After the propensity scores were created, teachers were sorted by grade level of teaching assignment and their propensity scores. Fellows and their closest matches in the comparison pool (up to two) in the same grade level were included in the analytic samples for the evaluation. Some Fellows lacked close matches within the limited comparison pool in the same grades and subjects within their districts and were thus excluded. The matching process was conducted separately for teachers in the first year and second year of teaching. (See Appendix D for more details on the matching process.)

The matching process produced the following three sets of analytic samples (see Appendix E for details):

- The sample to address the primary research question (impacts in the second year) consisted of 452 Fellows and 789 comparison teachers in the second year of teaching, combined across cohorts and districts. Fellows in this sample represented 78% of all Fellows from the 2010 through 2013 cohorts who remained in their districts' teacher records in their second year of teaching (see Figure 5).
- The sample to address the supplemental research question concerning teachers' first year of teaching consisted of 487 Fellows and 886 comparison teachers, combined across cohorts and districts. Fellows in this sample represented 85% of all Fellows from the 2011, 2012, and 2013 cohorts who were in their districts' teacher records in their first year of teaching (see Figure 5).
- The sample to address the supplemental, follow-up research question concerning teachers' third year of teaching consisted of 237 Fellows and 404 comparison teachers, combined across cohorts and districts. Fellows in this sample represented 78% of all Fellows from the 2010 through 2012 cohorts who remained in their districts' teacher records in their third year of teaching (see Figure 5).

Figure 5. Sample of Fellows for Analysis of Instructional Practice



Analyses Estimated Differences in Teachers' Instructional Practices

The differences in teacher outcomes between the Fellows and the comparison teachers were examined using regression analysis. This statistical technique allows for control of other teacher and school background characteristics. (See Appendix D for more details.) Because the initial data came from four district sites with different observation protocols and rubrics, we conducted parallel analyses for the sample of first-year teachers and the sample of second-year teachers within each district.[16]

[16] Estimation of impacts on teacher practices treated each district as a sub-sample with a separate analysis, resulting in district-specific impact estimates that were then averaged across districts to form an overall estimate. In contrast, estimation of impacts on student achievement used z-scores as a common achievement measure and combined samples across all districts.

This approach estimated the mean instructional practice scores for Fellows and comparison teachers after adjusting for the possible influence of teacher and school characteristics. The difference in the adjusted means on each outcome measure was converted into an effect size estimate that represented the magnitude of the difference in outcome scores in standard deviation units.

To estimate the difference between Fellows and comparison teachers across districts, we calculated a weighted average of the effect size estimates for the appropriate sites. Those districts with more precise estimates (larger sample sizes) were given more weight than those with less precise estimates (smaller sample sizes).

Methods for Comparing Teacher Retention

We compared the retention rates of Fellows relative to other new teachers in their second, third, and fourth years of teaching using longitudinal analysis. The data provided by the four districts included in the evaluation of teacher instructional practice also allowed examination of teacher retention for all new teachers in these districts. A longitudinal file was constructed for each cohort of new teachers in each district to track individual teachers from when they began teaching through academic year 2014–15. Retention was defined as a teacher (based on a unique random identification number) who was present in the district teacher records in each subsequent year. To calculate retention rates, the number of teachers within each group (Fellows or other new teachers) in each site who returned for the nth year and all previous years was divided by the number of teachers who began teaching in each cohort. Pearson's chi-squared tests were calculated to determine whether the observed differences were statistically significant.

No statistical matching methods were used for this analysis of teacher retention. All Fellows and other new teachers were included, with the exception of TFA teachers, who were excluded from the sample.

Impacts of the Teaching Fellows Program

Findings for Student Achievement

Matching Process Resulted in Comparable Groups of Students Taught by Fellows and Students Taught by Comparison Teachers

For each of the two analytic samples, we assessed whether the Fellow and comparison groups were equivalent on the pretest measure by checking whether the standardized average difference in prior student achievement between the two groups was less than or equal to the WWC threshold of 0.25 *SD*. Students of Fellows and students of comparison teachers in the matched samples had similar prior achievement, with a difference of about 0.03 *SD* for the second-year sample and a difference of approximately 0.00 *SD* for the first-year sample. The matching process created adequate balance in terms of baseline student achievement, which is consistent with the WWC standards for quasi-experimental designs (WWC, 2013).

We also compared baseline student and teacher demographic characteristics. In both analytic samples, all observed differences in demographic characteristics between students of Fellows and students of matched comparison teachers, as well as between Fellows and comparison teachers, were smaller than 0.25 *SD* (see Appendix E for more details).

Achievement Was Similar Between Students of Fellows and Those of Comparison Teachers

On average, students of Fellows performed similarly as did students of matched comparison teachers. With students pooled across districts and cohorts, the estimated difference in average achievement between the two groups was not significantly different from zero (Table 2).[17] This was the case for students of second-year teachers (0.017 *SD*, $p = .631$) and students of first-year teachers (–0.046 *SD*, $p = .078$).

[17] Retrospective power calculations show that a study with sample sizes and parameters similar to the current study would have 80% power to detect a minimum detectable effect size of 0.07 *SD* for the first year of the teaching sample and 0.10 *SD* for the second year of the teaching sample. These effect sizes are consistent with reported impacts for teacher preparation programs on student achievement (Clark et al., 2015; Decker et al., 2004; Xu et al., 2008).

Table 2. Estimated Difference in Average Student Achievement in the Analytic Samples

Outcome	Mean		Mean Difference	Standard Error	Effect Size	*p* Value
	Fellows	Comparison Group				
Second-Year Teachers Sample						
Outcome						
Student achievement (*z*-score)	-0.613	-0.627	0.014	0.029	0.017	.631
Sample Size						
Number of matched students	12,795	10,778				
Number of matched classes	587	970				
Number of matched teachers	303	693				
First-Year Teachers Sample						
Outcome						
Student achievement (*z*-score)	-0.372	-0.331	-0.041	0.023	-0.046	.078
Sample Size						
Number of matched students	18,826	18,273				
Number of matched classes	840	1,393				
Number of matched teachers	445	1,014				

Note. Means and difference in means are regression adjusted to account for the clustering of students within classrooms within teachers, and for students' achievement at baseline and the characteristics of students, classrooms, and teachers; district effects are weighted by the proportion of Fellows from each district in the analytic sample. The *p*-values are from a two-tailed test of the null hypothesis of equality of Fellow and comparison group means. Effect sizes were computed by dividing the difference by the pooled *SD* of the outcome for the Fellow and comparison groups (Hedges' *g*).

Source. AIR's analysis based on data provided by districts.

Achievement of Students of Fellows and Those of Comparison Teachers Was Similar Across Subgroups Defined by Districts, Subjects, Cohorts, and Grade Levels

Consistent with the findings on the full analysis samples, the estimated differences in academic achievement between students of Fellows and students of matched comparison teachers were not significantly different from zero across subgroups defined by districts, grade levels, subjects, or cohorts, both in the second-year sample and in the first-year sample (Figures 6 and 7).[18] As shown in the figures, the estimated differences varied in magnitude and direction across subgroups but were

[18] Because of the smaller sample sizes in these subgroups, subgroup analyses were statistically underpowered to detect differences in individual districts, grade levels, subject areas, or cohorts.

generally close to zero. Tests of the variation across subgroups were not significant, indicating that the differences across districts, subjects, cohorts, and grade levels were no larger than would be expected from chance.

Similar Findings Were Observed When Using Alternative Analytic Specifications

To examine the robustness of our findings to the analytic methods employed, we conducted a set of sensitivity analyses, including using alternative criteria for identifying samples (deletion or retention of cases with missing covariates or with multiple records); assigning alternative weighting schemes based on the proportion of students from each district in the analytic sample and the precision of the estimated district-specific differences; applying alternative specifications of analytic models based on which demographic characteristics were included[19]; and using an alternative two-level model of students nested within teachers.

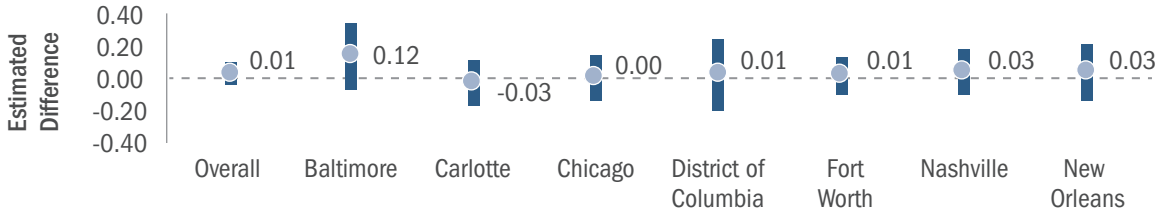
Findings from most of these alternative approaches were similar to findings from the main models, with no statistically significant differences in achievement between students of Fellows and students of comparison teachers. One exception was an alternative analysis that excluded cases with missing covariate values for the first-year of the teaching sample. This analysis resulted in a statistically significant negative effect that indicated that students of Fellows scored about 0.055 *SD* lower than students of matched comparison teachers ($p = .042$). Although this finding is significant, its estimated effect is similar in magnitude to that of the main analysis (effect size = –0.046 *SD*).[20] The main takeaway from these sensitivity analyses is that the magnitude and direction of the estimated differences were robust to alternative analytic approaches.

[19] Specifically, we fitted statistical models that excluded all demographic characteristics that were not statistically significant in the main analyses. Because neither of the two teacher demographic characteristics (gender and race) were statistically significant in the main analyses for both the first-year and second-year samples, these alternative models did not control for any teacher demographic characteristics.

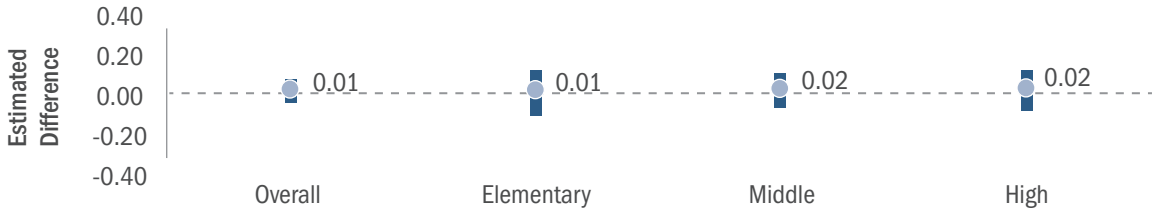
[20] About 3.8% of teachers in the first-year analytic sample had missing gender and/or race. These teachers came from three of the seven study districts. Among these teachers, students of Fellows tended to have higher achievement outcomes than students of comparison teachers in two districts. As a result, the estimated district-specific effects in these two districts became more negative compared to the estimated districts-specific effects from the main analysis. The change observed in the two districts resulted in an overall impact estimate with similar magnitude as that from the main analysis but with a *p* value that fell slightly below .05.

Figure 6. Estimated Differences in Average Student Achievement Between Fellows and Comparison Teachers in Their Second Year of Teaching, by Subgroup

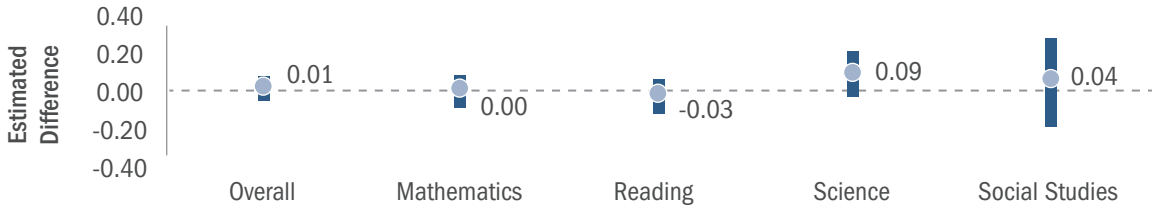
(a) By District



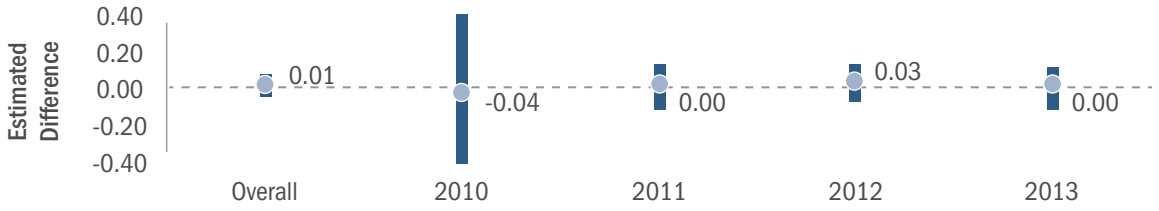
(b) By Grade Level



(c) By Subject



(d) By Cohort

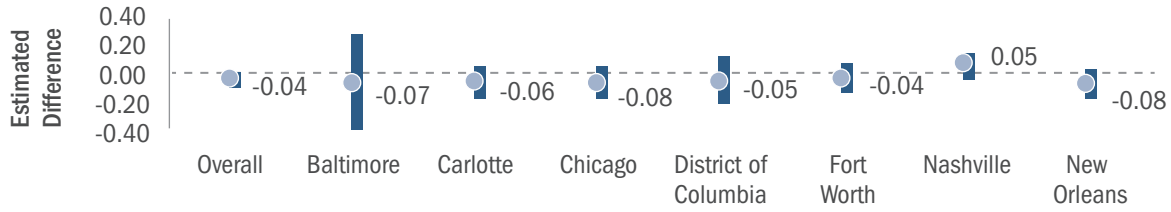


Note. Estimated mean differences are regression adjusted to account for the clustering of students within classrooms within teachers; students' achievement at baseline; and the characteristics of students, classrooms, and teachers. Performance is measured in z-scores that represent students' performance relative to the district average in a particular grade, subject, and cohort. The overall difference is an average of the district-specific estimates weighted by the proportion of Fellows from each district in the analytic sample. Each vertical line represents the 95% confidence interval around each estimated difference. A wider confidence interval means a larger standard error or uncertainty about the estimated difference. Confidence intervals that include zero indicate that the difference is not statistically significant at the .05 level. See Appendix E for sample sizes.

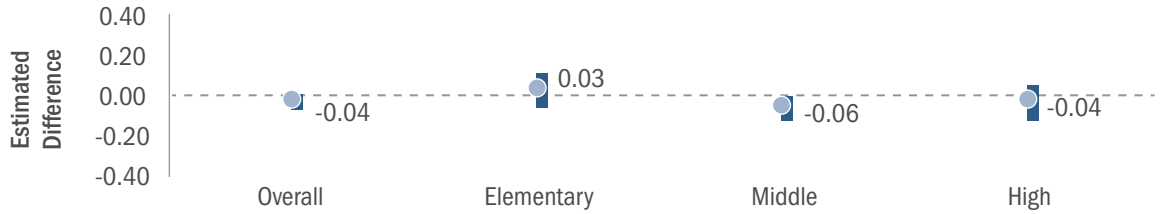
Source. AIR's analysis based on data provided by districts.

Figure 7. Estimated Differences in Average Student Achievement Between Fellows and Comparison Teachers in Their First Year of Teaching, by Subgroup

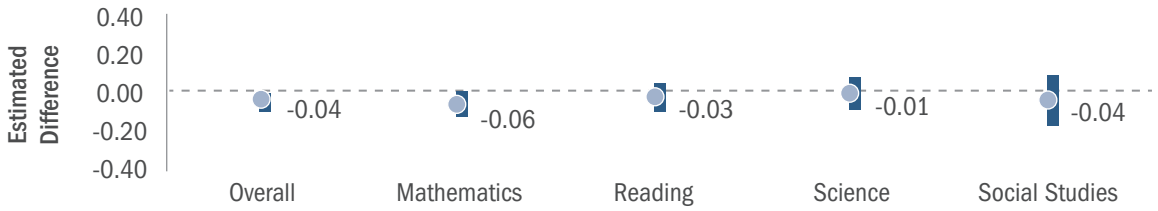
(a) By District



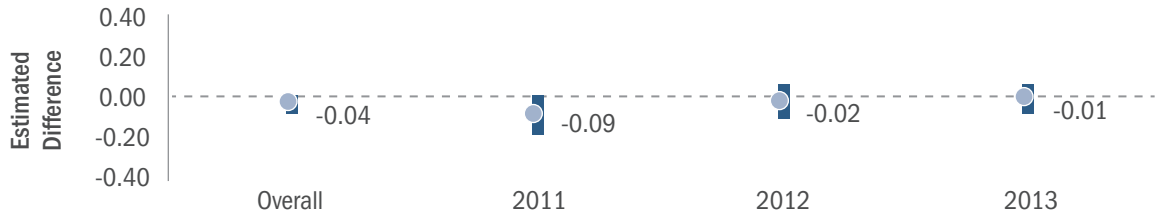
(b) By Grade Level



(c) By Subject



(d) By Cohort



Note. Estimated mean differences are regression adjusted to account for the clustering of students within classrooms within teachers; students' achievement at baseline; and the characteristics of students, classrooms, and teachers. Performance is measured in *z*-scores that represent students' performance relative to the district average in a particular grade, subject, and cohort. The overall difference is an average of the district-specific estimates weighted by the proportion of Fellows from each district in the analytic sample. Each vertical line represents the 95% confidence interval around each estimated difference. A wider confidence interval means a larger standard error or uncertainty about the estimated difference. Confidence intervals that include zero indicate that the difference is not statistically significant at the .05 level. See Appendix E for sample sizes.

Source. AIR's analysis based on data provided by districts.

Findings for Teacher Instructional Practice

Matching Process Resulted in Comparable Groups of Fellows and Comparison Teachers

For each district, we assessed whether Fellows and comparison teachers included in the analytic samples were equivalent in observed teacher and school background characteristics. Across districts, standardized differences in observed teacher and school characteristics were less than 0.25 *SD* in all years examined (see Appendix E for more details). These findings indicate that the matching process achieved adequate balance in terms of observed teacher and school characteristics.

Classroom Observation Scores Were Similar Between Fellows and Comparison Teachers

Fellows demonstrated similar instructional practice to comparison teachers in their second year of teaching (the main focus of this component of the evaluation), as measured by the classroom observations. When overall classroom observation scores were combined into a weighted average across districts, Fellows' scores were not significantly different from comparison teachers (g = 0.00,[21] p = .942; Table 3).[22]

The findings were consistent within the other years examined, with no significant differences between Fellows and comparison teachers in the first year of teaching (g = –0.10, p = .084) or the third year of teaching (g = 0.12, p = .131). Looking across years, the estimated difference in overall scores across districts moved from –0.10 *SD* in the first year of teaching to 0.12 *SD* in the third year of teaching. This finding suggests a relative improvement may occur for Fellows over time, though overlapping confidence intervals for the estimated differences across years indicate the changes over time were not statistically significant.

Generally, Fellows and comparison teachers also had similar scores in the Instruction and Classroom Environment domains of the observations. No differences were found in the Instruction domain for first-year (g = –0.09, p = .167), second-year (g = –0.02, p = .775), or third-year teachers (g = 0.11, p = .203). For Classroom Environment domain scores, first-year Fellows scored significantly lower than comparison teachers (g = –0.14, p = .026), weighted across districts; within-district analysis indicated that the difference in Classroom Environment scores was significant only in the District of Columbia. No significant differences were found in Classroom Environment scores for the more experienced teachers in the second year (g = 0.04, p = .509) or third year of teaching (g = 0.11, p = .199).

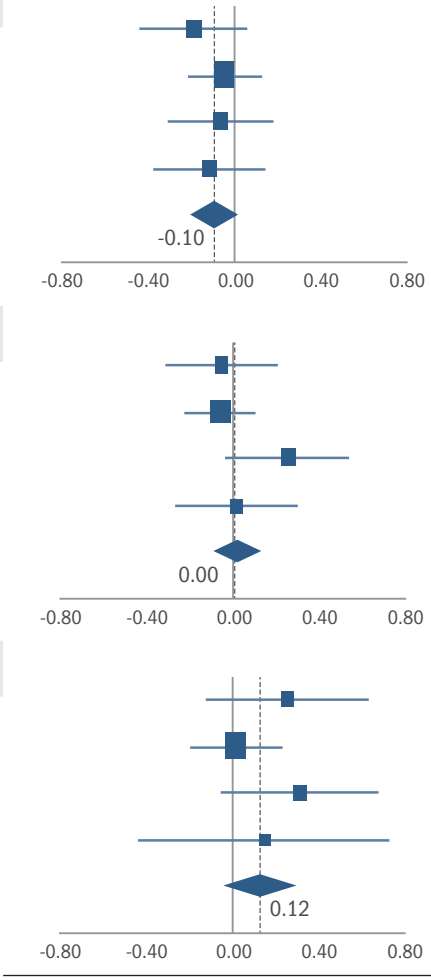
---

[21] g stands for Hedges' g, which represents standardized mean group difference.

[22] Retrospective power calculations show that a study with sample sizes and parameters similar to the current study would have 80% power to detect a minimum detectable effect size of 0.16 *SD* for the first year of the teaching sample, 0.17 *SD* for the second year of teaching sample, and 0.31 *SD* for the third year of the teaching sample.

Table 3. Differences in Overall Instructional Practice Scores Between Fellows and Comparison Teachers, Within and Across Districts

District	Difference in Means	Sample Size	Effect Size	*p* Value	Effect Size and 95% Confidence Interval (CI)
First-Year Teachers					
Chicago	-0.08	280	-0.19	.124	
District of Columbia	-0.03	573	-0.06	.498	
Nashville	-0.04	272	-0.07	.589	
New Orleans	-0.06	246	-0.12	.373	
Summary Effect		1,371	-0.10	.084	
Second-Year Teachers					
Chicago	-0.02	242	-0.05	.694	
District of Columbia	-0.02	590	-0.06	.478	
Nashville	0.13	199	0.26	.085	
New Orleans	0.01	206	0.02	.892	
Summary Effect		1,237	0.00	.942	
Third-Year Teachers					
Chicago	0.08	118	0.26	.186	
District of Columbia	0.01	355	0.02	.886	
Nashville	0.18	121	0.31	.098	
New Orleans	0.08	47	0.15	.892	
Summary Effect		641	0.12	.131	



Note. The means are regression adjusted to account for the differences in teacher and school characteristics. The *p* values are from a two-tailed test of the null hypothesis of equality of means for Fellows and comparison teachers. In the plot, each box in a line indicates the estimate for each district-specific effect: Size of box = the weight given to the district-specific effect; Width of line = 95% CI for the district-specific effect. Diamond indicates the overall summary effect: Edge of diamond = 95% CI for the summary effect.

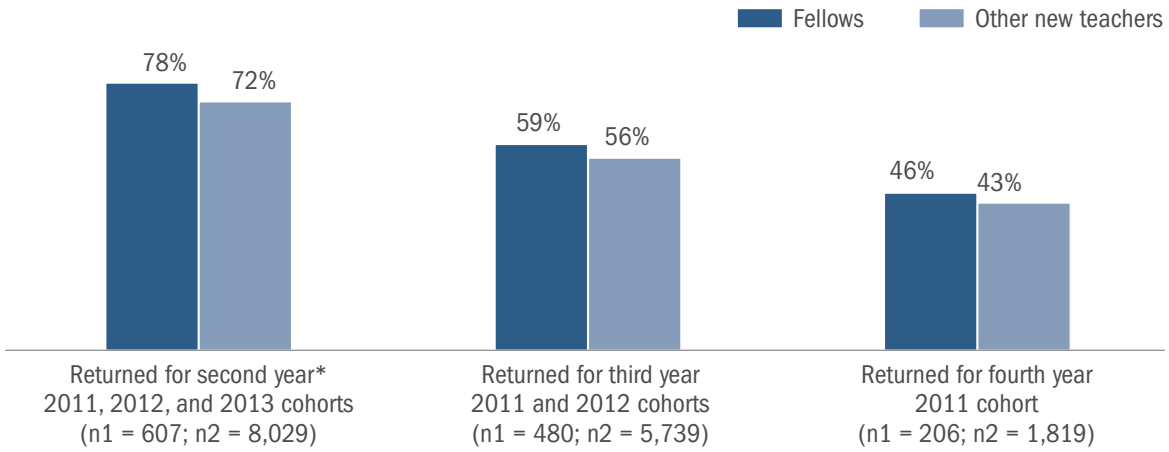
Source. AIR's analysis based on data provided by Chicago Public Schools, District of Columbia Public Schools, Metropolitan Nashville Public Schools, and the Louisiana Department of Education.

Findings for Teacher Retention

Averaged across sites and cohorts, the retention rate for Fellows in the second year was six percentage points higher than for other new teachers in their districts, a statistically significant difference (Figure 8). By cohort, Fellows in the 2011 and 2012 cohorts had significantly higher retention rates in the second year; no significant differences were found for the 2013 cohort (Figure 9). By district site, Fellows in Nashville demonstrated significantly higher retention rates than other teachers in the second and third years; none of the differences within other districts were statistically significant (Figure 10).

The cohorts of Fellows included those individuals who ultimately did not pass TNTP's Assessment of Classroom Effectiveness at the end of the first year of teaching, approximately 8% of Fellows per year according to TNTP. These nonpassing teachers represent a form of intentional attrition by TNTP, through removal of the lowest performers from the cohort (e.g., not recommending them for certification) prior to the second year of teaching. If these teachers had been excluded from the cohort analysis, the positive difference in retention for Fellows would have likely been larger. However, we did not have information in district data about ACE scores for enough Fellows to allow for exclusion of nonpassing Fellows.

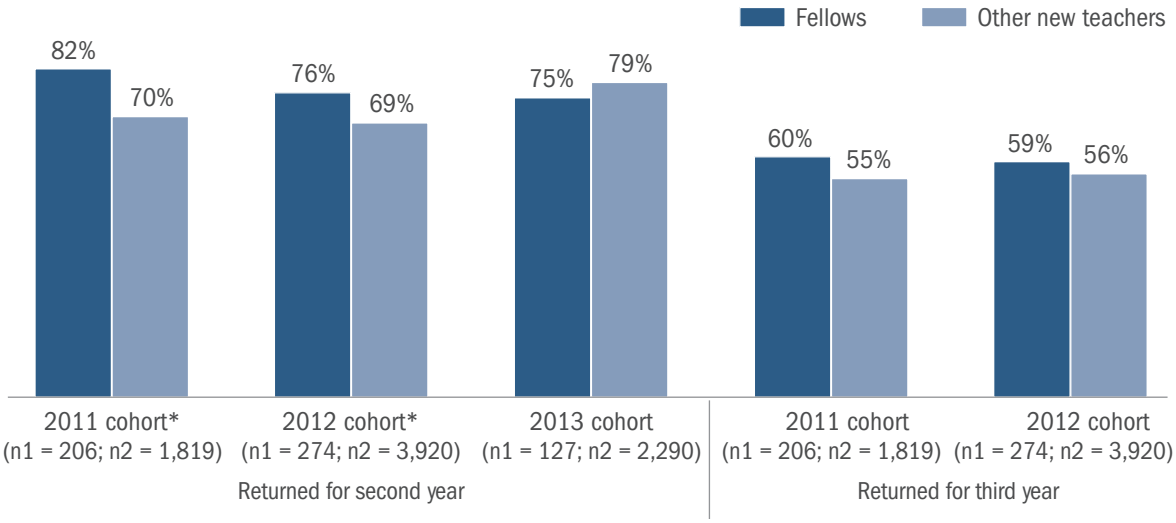
Figure 8. Overall Retention Between Fellows and Other New Teachers, All Sites and Cohorts



Note. n1 = number of Fellows; n2 = number of other new teachers; *The difference in retention rates between Fellows and other new teachers is statistically significant at the .05 level.

Source. AIR's analysis is based on data provided by districts or state agencies.

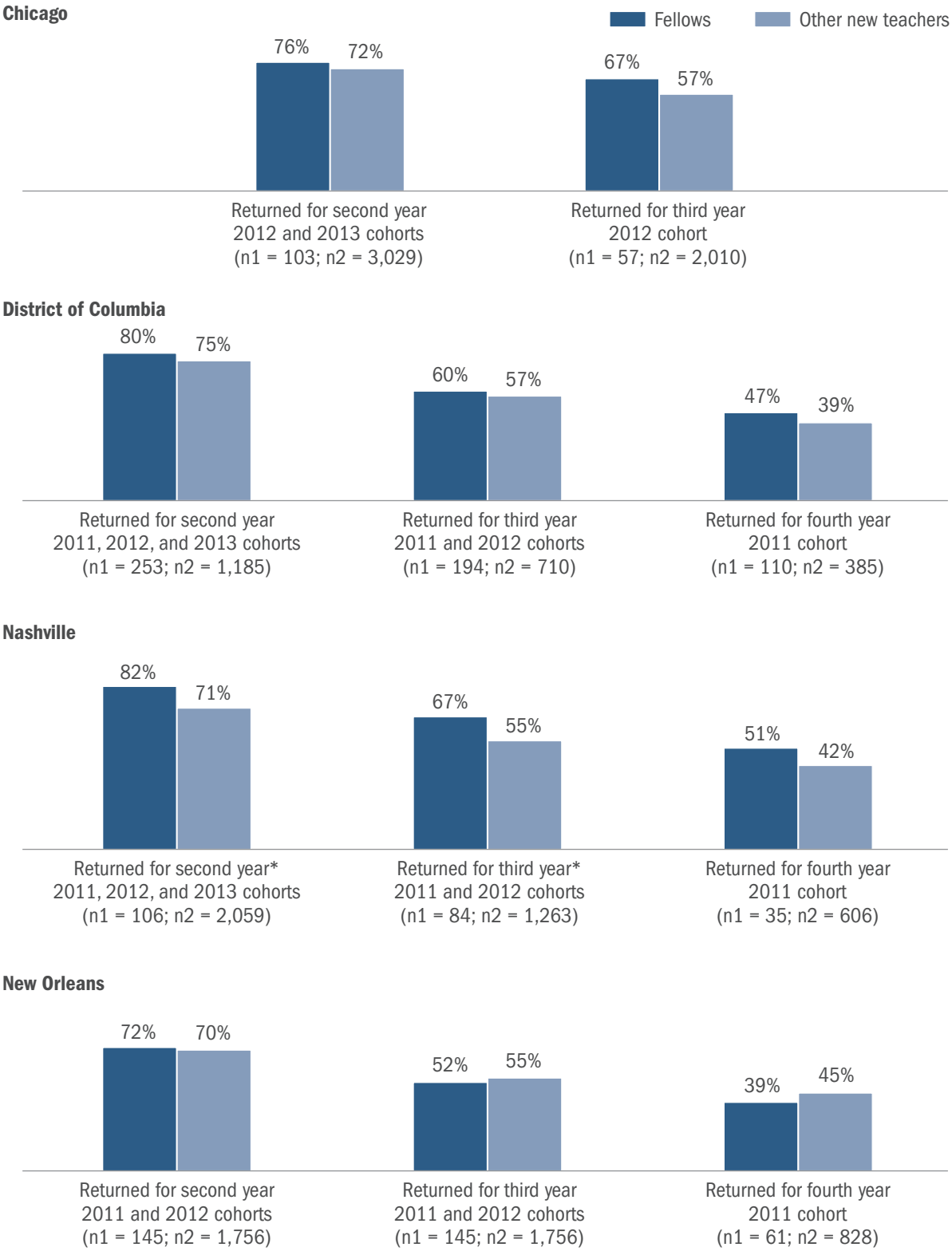
Figure 9. Across-Site Retention Rates Between Fellows and Other New Teachers, by Cohort



Note. n1 = number of Fellows; n2 = number of other new teachers; *The difference in retention rates between Fellows and other new teachers is statistically significant at the .05 level.

Source. AIR's analysis is based on data provided by districts or state agencies.

Figure 10. Across-Cohort Retention Rates Between Fellows and Other New Teachers, by District



Note. n1 = number of Fellows; n2 = number of other new teachers; *The difference in retention rates between Fellows and other new teachers is statistically significant at the .05 level.

Source. AIR's analysis is based on data provided by districts or state agencies.

Conclusions

Across all participating districts, TNTP implemented the Teaching Fellows program with fidelity to the program design under the i3 grant. The study found that students taught by Fellows performed similarly to students taught by comparable teachers prepared through other programs and that Fellows and comparable teachers demonstrated similarly effective classroom instructional practice.

In the districts and Teaching Fellows cohorts examined, TNTP recommended certification for approximately 1,200 new teachers, who were retained into the second year of teaching at higher rates than other new teachers in their districts. These Fellows were found to have similar performance even though their training period was shorter than is typically provided through traditional teacher preparation programs. Considering the persistent needs for qualified teachers in urban districts, the study indicates that the Teaching Fellows program recruited and trained qualified teachers and provided a viable pathway for new teachers in the partner districts. However, the study's findings suggest that TNTP fell short of its goal to produce a cadre of new teachers who were more effective than other new teachers hired to fill vacancies in the study districts.

Two factors may have contributed to why the Fellows program as evaluated here did not meet TNTP's objective of providing teachers who outperformed their district peers. First, during the evaluation period, districts and states made considerable changes to their teacher evaluation systems, including incorporating student test scores and multiple measures of teacher instructional performance (Hull, 2013). Some of these changes mirrored the expectations TNTP used for its Fellows. As a result, the differentiation between the TNTP Fellow experience and "business as usual" may have become smaller over time, reducing the program contrast evaluated in this study. Second, by design, programs such as TNTP increase the overall applicant pool for teaching positions in a district. By doing so, TNTP may have indirectly improved the quality of the pool of non-TNTP teachers by allowing districts and school leaders to be more selective in filling positions.

Compared with prior research on the Teaching Fellows program, this study provides a large-scale analysis of Fellows and their students across multiple sites, grade levels, and subjects, representing a breadth of conditions in which Fellows taught. The study used comparison groups of teachers with the same level of experience and is the first we are aware of to examine the impacts of the Teaching Fellows program using an instructional practice measure. Our findings are largely consistent with prior studies on the Teaching Fellows program, which have reported at most small impacts on student test performance (typically with effect sizes of 0.05 *SD* or smaller) when comparing Fellows to other teachers (Boyd et al., 2006; Clark et al., 2013; Kane et al., 2006). Some prior studies found that Fellows improved performance relative to comparison teachers after the first year of teaching (Boyd et al., 2006; Kane et al., 2006). Our findings on teachers' classroom observation scores suggest a similar trend. Further research more focused on outcomes for Fellows' classroom instruction and teaching practice may be informative to TNTP's efforts to improve program design and implementation.

Limitations

The methods used for this study have several limitations. First, although the propensity score matching process resulted in similar comparison groups based on the available data, it is possible that the groups differed systematically on measures that were not available for the analysis and that might be related to the selection of teachers into the TNTP program and the outcomes measured. Fellows chose to participate in the Teaching Fellows program, and TNTP may attract candidates whose background profiles are meaningfully different from their peers. The available data for this study may not fully account for selection into the Teaching Fellows program. If Fellows were different from comparison teachers in characteristics not accounted for in the study, the reported findings cannot be isolated from these characteristics.

Second, only Fellows in grades and subjects assessed by the state tests and with students with observed pretest and posttest scores were included in the pool of teachers for matching used to estimate impacts on student achievement. As stated, the analytic sample for second-year teachers represented only 29% of all Fellows in the district records in their second year of teaching, and the analytic sample for first-year teachers made up 37% of all Fellows in the district records in their first year of teaching. These findings would not necessarily generalize to all Fellows hired by the district.

Lastly, based on the existing research (Kane & Staiger, 2012; Wayne et al., 2016), teacher observation data could have limited reliability. Although the observation protocols and processes in the four districts included in the analyses of instructional practices all met the criteria set for this evaluation, the observation data received from districts show various degrees of consistency across observers and observation cycles (see Appendix B). The inherent variability across observations, for individual teachers, might make it more difficult to observe differences between teachers. Because a single observation by a single observer is generally considered an unreliable estimate of a teacher's instructional practice (Ho & Kane, 2013; Kane & Staiger, 2012), AIR used average scores from multiple observations as the outcome measures.

References

Aaronson, D., Barrow, L., & Sanders, W. (2003). *Teachers and student achievement in the Chicago public high schools* [Working Paper No. 2002-28]. Chicago, IL: Federal Reserve Bank of Chicago.

Anderson, J. (2013, March 30). Curious grade for teachers: Nearly all pass. *The New York Times,* A1.

Barrett, N., & Harris, D. (2015). *Significant changes in the New Orleans teacher workforce.* New Orleans, LA: Education Research Alliance for New Orleans. Retrieved from http://educationresearchalliancenola.org/files/publications/ERA-Policy-Brief-Changes-in-the-New-Orleans-Teacher-Workforce.pdf

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* West Sussex, UK: Wiley.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy, 1*(2), 176–216.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* [NBER Working Paper 17699]. Cambridge, MA: National Bureau of Economic Research.

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). *The effectiveness of secondary math teachers from Teach For America and the Teaching Fellows programs* (NCEE 2013-4015). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Clark, M. A., Isenberg, E., Liu, A. Y., Makowsky, L., & Zukiewicz, M. (2015). *Impacts of the Teach For America Investing in Innovation Scale-Up.* Princeton, NJ: Mathematica Policy Research.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673–682.

Cohen, J. & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378–387

Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). *An evaluation of teachers trained through different routes to certification, final report* (NCEE 2009-4043). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Coopersmith, J. (2009). *Characteristics of public, private, and Bureau of Indian Education elementary and secondary school teachers in the United States: Results from the 2007–08 schools and staffing survey* (NCES 2009-324). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

DeAngelis, K. J., Presley, J. B., & White, B. R. (2005). *The distribution of teacher quality in Illinois* (IERC 2005-1). Edwardsville, IL: Illinois Education Research Council.

Feistritzer, C. E. (2011). *Profiles of teachers in the U.S. 2011.* Washington, DC: National Center for Education Information.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis, 37*(2), 224–242.

Goldhaber, D. (2002). The mystery of good teaching. *Education Next, 2*(1), 50–55.

Gordon, R., Kane, T., & Staiger, D. (2006). *Identifying effective teachers using performance on the job* [Hamilton Project White Paper]. Washington, DC: Brookings Institution.

Greenberg, J., McKee, A., & Walsh, K. (2013). *Teacher prep review: A review of the nation's teacher preparation programs.* Washington, DC: National Council on Teacher Quality. Retrieved from http://www.nctq.org/dmsStage/Teacher_Prep_Review_2013_Report

Grossman, P., & Loeb, S. (2010). Learning from multiple routes. *Education Leadership, 67*(8), 22–27.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1998). *Teachers, schools, and academic achievement* [NBER Working Paper 6691]. Cambridge, MA: National Bureau of Economic Research.

Harris, D. N., & Sass, T. R. (2007). *Teacher training, teacher quality and student achievement. Working paper 3.* Washington, DC: Urban Institute, Center for the Analysis of Longitudinal Data in Education Research.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (MET Project Research Paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://eric.ed.gov/?id=ED540957

Hull, J. (2013). *Trends in teacher evaluation: How states are measuring teacher performance.* Alexandria, VA: Center for Public Education.

Ingersoll, R. (2003). *Is there really a teacher shortage?* Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Ingersoll, R., & May, H. (2012). The magnitude, destinations, and determinants of mathematics and science teacher turnover. *Educational Evaluation and Policy Analysis, 34*(4), 435–464.

Isenberg, E., Max, J., Gleason, P., Johnson, M., Deutsch, J., & Hansen, M. (2016). *Do low-income students have equal access to effective teachers? Evidence from 26 districts* (NCEE 2017-4008). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Jacobs, B. (2007). The challenges of staffing urban schools with effective teachers. *The Future of Children, 17*(1), 129–153.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City* (Working Paper 12155). Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains* (MET Project research paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://files.eric.ed.gov/fulltext/ED540960.pdf

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Education Evaluation and Policy Analysis, 24*(1), 37–62.

Lemov, D. (2010). *Teach like a champion: 49 techniques that put students on the path to college* (K–12). San Francisco, CA: Jossey-Bass.

McKibbin, M. D. (1998). *Voices and views: Perspectives on California's teaching and internship programs.* Sacramento, CA: Commission on Teacher Credentialing.

Miller, R., & Chait, R. (2008). *Teacher turnover, tenure policies, and the distribution of teacher quality: Can high-poverty schools catch a break?* Washington, DC: Center for American Progress.

Mulhern, J., Grogan, E., & Wexler, D. (2013). *Leap year: Assessing and supporting effective first-year teachers.* Brooklyn, NY: TNTP. Retrieved from http://tntp.org/assets/documents/TNTP_LeapYear_2013.pdf

Noell, G. H., Gansle, K. A., Patt, R. M., & Schafer, M. J. (2009). *Value added assessment of teacher preparation in Louisiana: 2005–2006 to 2007–2008.* Baton Rouge, LA: Louisiana State University, Department of Psychology.

Ng, J., & Peter, L. (2010). Should I stay or should I go? Examining the career choices of alternatively licensed teachers in urban schools. *Urban Review, 42*(2): 207–227.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247–252.

Sanders W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement* [Research progress report]. Knoxville: University of Tennessee, Value-Added Research and Assessment Center.

Shen, J. (1999). Alternative certification: Math and science teachers. *Educational Horizons, 78*(1), 44–48.

TNTP. (2014). *Preparing teachers for a Fast Start: A new approach to beginning teacher training.* Retrieved from http://www.teachingworks.org/images/files/Preparing_Teachers_for_a_Fast_Start.pdf

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service. (2016). *The state of racial diversity in the educator workforce.* Washington, DC: Author. Retrieved from https://www2.ed.gov/rschstat/eval/highered/racial-diversity/state-racial-diversity-workforce.pdf

U.S. Department of Education, Office of Postsecondary Education. (2015). *Teacher shortage areas nationwide listing 1990–1991 through 2015–2016.* Washington, DC: Government Printing Office. Retrieved from http://www2.ed.gov/about/offices/list/ope/pol/tsa.pdf

Wayne, A. J., Garet, M. S., Brown, S., Rickles, J., Song, M., & Manzeske, D. (2016). *Early implementation findings from a study of teacher and principal performance measurement and feedback: Year 1 report* (NCEE 2017-4004). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

What Works Clearinghouse. (2013). *What Works Clearinghouse: Procedures and standards handbook* (Version 3.0). Washington, DC: Author. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts.* Washington, DC: Brown Center on Education Policy and Brookings Institute.

Wilson, S. M., Floden, R., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations.* Seattle, WA: University of Washington, Center for the Study of Teaching and Policy.

Appendix A. Description of the Teaching Fellows Program

Program Components

The Teaching Fellows program implemented under the i3 grant was structured according to four primary components: recruitment and selection, preservice training, inservice training, and the Assessment of Classroom Effectiveness (ACE). These four components were maintained throughout the implementation of the program supported by the i3 grant. TNTP also adjusted the program model during the grant period. Locally, TNTP site staff adjusted the operation of the program based on district hiring needs, progress in recruitment, and other location-specific conditions. Centrally, TNTP introduced substantive changes to the program to more rapidly build instructional skills among the Fellows, based on inservice teacher performance data and feedback from the Fellows through periodic surveys. This section describes the program structure and highlights changes in the program during the grant period.

Recruitment and Selection

Each district site was expected to set its recruitment targets for high-need subjects and grades using multiple strategies to recruit applicants, such as marketing through the Internet and print media, developing connections with community organizations, and making community and campus presentations. TNTP staff screened applicants by reviewing their application materials, analyzing a written hiring exercise (or, in some cases, the submission of a work sample), and conducting telephone and in-person interviews. This multipronged process was designed to select Fellow applicants who showed the most promise for becoming effective teachers.

According to the TNTP staff interviewed by AIR, the Teaching Fellows program implemented under the i3 grant met the target numbers of Fellows placed in all grant years. Some sites (e.g., Chicago and Fort Worth) had slower than expected starts because of local budget and staffing considerations, but others (especially New Orleans) demonstrated higher than expected numbers of job openings. Local TNTP staff adjusted their recruitment and initial selection procedures to meet the recruitment targets TNTP agreed to at the outset of the i3 grant. TNTP staff recalled initial concerns about having to weaken recruitment standards to meet recruitment targets, but they observed across time that the Fellows admitted under the revised standards were "surprisingly successful" in the preservice program and in their first year of teaching. Thus, the revised recruitment standards were considered adequate.

Preservice Training

The Fellows selected into the program were expected to participate in a 5- to 8-week summer preservice training course provided by local TNTP sites. Prior to the summer training, the Fellows were required to complete four self-guided online modules on foundational knowledge about the teaching profession. The summer training involved approximately 75 hours of coursework centered on field experiences. During the training, the Fellows studied foundational skills and proven teaching techniques, rehearsed these skills, received coaching on them, and practiced them with peers and students in summer

school classrooms. Preservice training instructors were responsible for the development of the Fellows through skill-building sessions. Teacher development coaches were responsible for the development of the Fellows through observations, debriefing sessions, responsive coaching sessions, and evaluations.

While completing their preservice training, the Fellows were expected to apply for teaching positions in high-need schools within the partner districts. TNTP staff and instructors evaluated the Fellows with an end-of-training screening to identify those who showed the potential to be highly effective teachers. Participants who passed the end-of-training screening were allowed to enter the classroom and commence inservice training. Participants who did not meet the screening criteria were not permitted to continue in the program.

Changes Made to the Preservice Component During the Grant Period

For the 2010 and 2011 cohorts of Fellows supported by the i3 grant, the preservice training used in TNTP's Teach for Student Achievement curriculum covered two broad domains: instructional planning and delivery plus classroom management and culture. Six training modules were used to address these domains. Each module required three to six sessions, and for each session, the objectives were specified in advance, as were the activities designed to actualize the objectives. Each session also was linked to teacher competencies, which were presumed to be linked to teaching effectiveness. The Fellow Advisors, who were experienced teachers from the partner district, led the training sessions.

For the 2012 cohort, TNTP introduced substantial changes to the program. TNTP found that many Fellows in their first year of teaching did not improve their core teaching skills substantially during their first year in the classroom. In response, TNTP incorporated the Fast Start training approach. Fast Start was designed to more quickly provide Fellows with the skills that teachers need to be effective early in their careers based on three core principles: focus, practice, and feedback (TNTP, 2014).

The Fast Start curriculum focuses on the development of the following skills: delivering lessons clearly, maintaining high academic expectations, maintaining high behavioral expectations, and maximizing instructional time. To build the Fast Start curriculum, TNTP worked closely with Uncommon Schools, an organization that developed a series of 17 teaching techniques that were featured in *Teach Like a Champion* (Lemov, 2010). Initially, in 2012, preservice training focused on all 49 of the *Teach Like a Champion* techniques. Later, TNTP identified that 19 techniques aligned with the four Fast Start skills, and these 19 techniques were emphasized in the preservice training. Fellows needed to demonstrate proficiency in 12 prioritized techniques during preservice training to earn eligibility to begin teaching in the fall. Their proficiency levels on those techniques were assessed in classroom observations and practice setting, as well as through submitted classroom artifacts. Four of the 19 techniques (100%, What to Do, Strong Voice, and Positive Framing) were found to have a strong correlation with Fellows' performance during the school year.[23] These four skills were called Anchor Techniques and were assessed by coaches during field experience. Under the Fast Start approach, Fellows spent 26 hours in intensive, hands-on practice prior to enacting the skills in actual classrooms.

[23] More information on Uncommon Schools' Taxonomy of Effective Teaching Practices can be found at http://uncommonschools.org/our-approach/teach-like-a-champion.

The Fast Start approach also incorporated coaching as an integral part of the preservice training. TNTP designed the coach role to provide immediate and specific instructional feedback while the Fellows were practice teaching, not just after a lesson, with feedback focused narrowly on one or two aspects of instruction. The Fellows were supposed to leave each coaching session with a specific and manageable list of things to work on for the next time.

The Fast Start approach was implemented fully in three sites (District of Columbia, Nashville, and New Orleans) and implemented partially in three sites (Baltimore, Charlotte-Mecklenburg Schools, and the Fort Worth Independent School District) for the 2012 cohort of Fellows. All sites fully implemented the Fast Start approach with the 2013 cohort, which was revised to have a more specific focus on the four key skills. For the purposes of this evaluation, the 2010 and 2011 cohorts of Fellows were not affected by the introduction of the Fast Start approach. The 2012 cohort was the first to be trained under Fast Start, based on partial implementation. The 2013 cohort was the first to experience the fully implemented Fast Start curriculum and training.

Inservice Training

During their first year working as full-time teachers, the Fellows took part in inservice training through the TNTP Academy and completed licensure coursework. The centerpiece of the academy coursework was a series of biweekly professional development seminars specific to participants' subject areas and grade levels. The seminars focused on content-specific instruction and teaching strategies and emphasized practical application of the teaching skills and techniques that participants studied during preservice training. Each seminar series included at least 16 sessions, and each seminar session lasted approximately 3 hours. To continue in the program, participants were expected to attend at least 14 of the 16 sessions.

Changes to the Inservice Component During the Grant Period

For the 2010 and 2011 cohorts of Fellows, the inservice professional development was centered on TNTP's Teaching for Results curriculum. The Fellows needed to attend 16 seminar sessions, and each seminar series was specific to teachers of particular subjects and grade levels. Each seminar session lasted 3 hours. Teaching for Results was intended to teach participants how to adjust their instruction to increase the academic achievement of students who were below grade level while also bolstering the academic achievement of students who were meeting grade-level expectations. The seminar topics and activities were intended to help participants better understand the content that students must master, communicate that content so that students can learn and apply it, administer assessment tools, and use assessment data to inform instruction. The seminar series for all content areas focused on three general competencies: (a) understanding the content domain that students must master, (b) using assessment tools to inform instruction, and (c) understanding instructional strategies that have been shown to be effective within the content domain.

The introduction of the Fast Start training approach in 2012 incorporated coaches into the TNTP Academy. Coaches were experienced local teachers selected by TNTP staff. They were expected to observe participants regularly in classrooms, provide real-time feedback on strategies that could be applied in future lessons, and place participants in small skill-building sessions to address Fellows' high-priority developmental needs.

Assessment of Classroom Effectiveness

Another key component of the Teaching Fellows program implemented under the i3 grant was TNTP's ACE. The ACE-based appraisal of Fellows' performance occurred at the end of the inservice training, and their appraisal scores determine whether Fellows were recommended for certification. TNTP designed ACE to include the following measures:

- Teachers' completion of inservice training
- Principals' evaluation of Fellows
- Classroom observations and student learning outcomes based on student surveys and achievement data wherever possible

Fellows who earned a passing score and successfully completed all program and state regulatory requirements were recommended for certification. Fellows who did not earn a passing score were not recommended for licensure by TNTP. If Fellows fell short of the passing score but provided evidence suggesting that they were on a trajectory to meet TNTP's standards, they were granted an extension year to continue building their skills.[24] Fellows who did not receive passing scores had the option to appeal the decision by sharing additional evidence and documentation about their performance.

Numbers of Fellows in the Cohorts Examined

The Teaching Fellows program enrolled 3,251 Fellows across the seven sites and cohorts included in this evaluation. Among the 1,642 who completed preservice training and moved into inservice training, 1,195 (or 73%) ultimately passed ACE and were recommended for certification. Some Fellows either dropped out at each step voluntarily or were removed from the program for failing the program requirements (which is part of the program design). Table A1 shows the distribution of Fellows who reached each program milestone by site.

[24] TNTP set the cutoff scores for passing. For the 2011 cohort, Fellows could earn up to 10 points—5 points for classroom observations and student learning outcomes, 3 points for principal ratings, and 2 points for program completion. Fellows who earned seven points or more were considered ready for certification. Fellows who scored between four and six points were placed on an extension plan, which allowed them another year in the program to continue building their skills. Fellows who scored three points or fewer were removed from the program (Mulhern, Grogan, & Wexler, 2013). For the 2012 and 2013 cohorts, TNTP adjusted the scoring and calculated the scores with a weighted formula. In this formula, Fellows could earn up to five points on each measure (classroom observations, student survey, principal rating, and student achievement). The weight assigned to each measure varied based on the number and mix of measures available for consideration. The final score was calculated by multiplying the points that Fellows earned for each measure by its relative weight and then adding those figures together. Fellows who scored 2.75 or higher on a 5-point scale were considered ready for certification. Fellows who scored between 2.50 and 2.74 points were placed on an extension plan, and those who scored fewer than 2.50 points were removed from the program without earning certification (Mulhern et al., 2013).

Table A1. Number of Fellows by Site According to TNTP Records

	Baltimore	Charlotte	Chicago	District of Columbia	Fort Worth	Nashville	New Orleans	All Seven Sites
Enrolled in the program	387	592	409	431	420	490	522	3,251
Started preservice training	251	396	248	480	248	335	398	2,356
Completed preservice training	220	269	210	367	210	248	301	1,825
Moved into inservice training	225	262	117	354	169	224	291	1,642
Completed inservice training	161	228	107	303	146	198	234	1,377
Recommended for certification	140	197	100	251	129	173	205	1,195

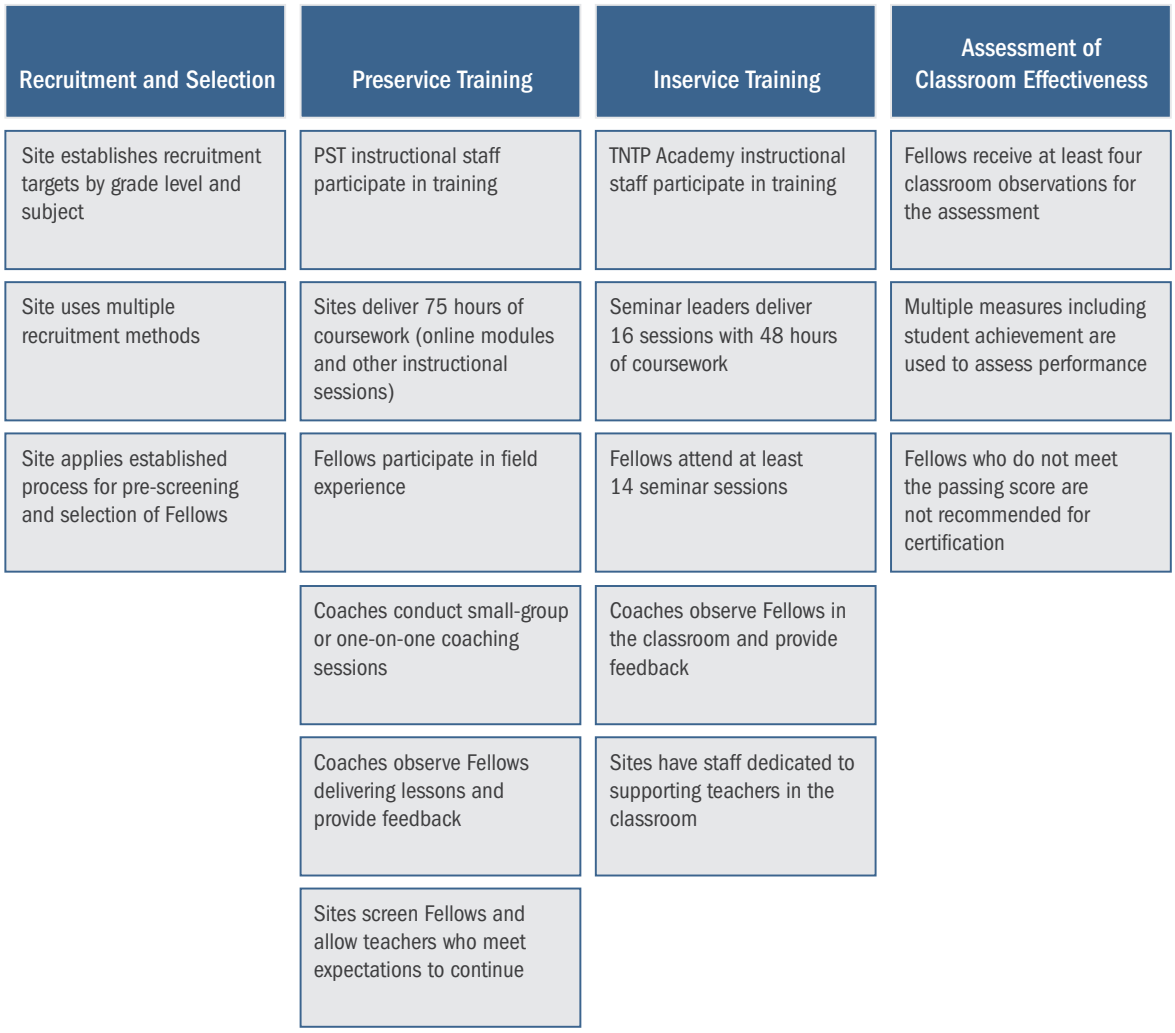
Note. AIR's analysis is based on program records provided by TNTP. TNTP provided site-level counts for the number of Fellows who enrolled in the program and individual-level records for the other milestones presented in this table. The table summarizes data for Fellows in the 2011, 2012, and 2013 cohorts in each district. Comparable data were not provided for the 2010 pilot cohort.

Fidelity of Implementation

AIR used two types of measures to examine fidelity of implementation: (a) site-level indicators (the extent to which the core components of the program were delivered as originally intended by TNTP in the seven sites) and (b) teacher-level indicators (the extent to which the Fellows perceived the training they received as adequate). Evidence for the former was based on program delivery and participation records provided by TNTP's central team and individual sites. Evidence for the latter was based on aggregated Fellow responses to TNTP-administered surveys.

Using the logic model underlying the program as a guide (Figure A1) and input from TNTP staff, we identified at least one observable indicator for each key program activity and thresholds for adequate implementation for each indicator. Indicators for the recruitment and selection component focused on the process of screening candidates, training interviewers, and the application-to-selection ratio. The preservice and inservice training indicators focused on Fellows' hours of participation in program activities and Fellows' ratings of the relevance and quality of the support they received from coaches and trainers. The indicators for ACE focused on implementation of the assessment procedures and Fellows' understanding of the rating and assessment process. We included 56 indicators across the four program components to measure fidelity of implementation.

Figure A1. Key Activities Measured for Fidelity of Implementation



The fidelity of implementation analysis summarized the results across these indicators in the four program components: recruitment and selection, preservice training, inservice training, and ACE. Each indicator had a designated threshold for adequate implementation. An indicator was scored as "1" if the threshold was observed and "0" if not. For each teacher-level indicator, sites were assigned a score equal to the proportion of Fellow responses that reflected the perceived presence of the indicator. The implementation scores (0 or 1) of all site-level indicators were averaged to obtain an aggregated implementation score for each site. Similarly, the implementation scores for all teacher-level indicators, each consisting of a proportion, also were averaged. The aggregated scores for the site-level and teacher-level indicators were averaged together within the appropriate program component, with a predetermined designation of .80 on average representing adequate implementation for a program component. (The maximum possible average was 1.0.) Fidelity scores were calculated for each cohort and then aggregated across the cohorts in each site.

TNTP's implementation of the program, as measured by these indicators and methods, met the thresholds established for this evaluation. Table A2 summarizes the results by both program component and site. All district sites met the implementation score benchmark for each program component, with aggregated scores across cohorts ranging from 0.83 to 0.98. The overall fidelity of implementation scores for each program component, which represents the average of sites' (aggregated) program component scores, were generally high, ranging from 0.88 to 0.95. Sites tended to score lower on the teacher-level indicators than on the site-level indicators, particularly for the inservice teaching and the ACE components. The lower scores suggested that TNTP should consider feedback from Fellows in providing guidance and oversight to site staff and coaches involved in program implementation and explore why the survey responses suggest weak or limited perceived adequacy of implementation.

Table A2. Fidelity of Implementation of Program Components Across Cohorts (2011, 2012, and 2013), by Site and Programwide

	Baltimore	Charlotte	Chicago	District of Columbia	Fort Worth	Nashville	New Orleans	Average Across Sites
Recruitment and Selection								
Implementation score for the component (site-level indicators only)	.95	.92	.95	.97	.93	.93	.97	.95
Preservice Training								
Implementation score for site-level indicators across cohorts	1.00	.97	1.00	.96	1.00	1.00	.97	–
Implementation score for teacher-level indicators across cohorts	.92	.92	.95	.91	.93	.85	.92	–
Implementation score for the component (average of site- and teacher-level scores)	.96	.95	.98	.93	.97	.93	.95	.95
Inservice Training								
Implementation score for site-level indicators across cohorts	1.00	1.00	.93	1.00	1.00	1.00	.93	–
Implementation score for teacher-level indicators across cohorts	.67	.71	.73	.80	.85	.86	.87	–
Implementation score for the component (average of site- and teacher-level scores)	.84	.85	.83	.90	.92	.93	.90	.88
ACE								
Implementation score for site-level indicators across cohorts	1.00	1.00	1.00	1.00	1.00	1.00	.87	–
Implementation score for teacher-level indicators across cohorts	.72	.80	.87	.76	.82	.80	.79	–
Implementation score for the component (average of site- and teacher-level scores)	.86	.90	.94	.88	.91	.90	.83	.88

Note. For each component, the implementation scores for individual sites were averaged to obtain an implementation score across sites, weighted by the average number of Fellows in the 2011, 2012, and 2013 cohorts in each site. Average implementation scores of .80 or greater were considered as meeting expectations for adequate implementation. Cells marked with a dash indicate there is no data, because averages across sites were calculated at the level of each component.

Appendix B. Instructional Practice Measures

To examine whether the Fellows demonstrated instructional practices that were more effective, we used teachers' scores from the classroom observation components of the state or district teacher evaluation systems as the outcome measures. In all four districts included in the analysis of teacher instructional practice outcomes (Chicago, District of Columbia, Nashville, New Orleans), teachers received multiple observations (by the same or different observers) each year and generally were rated on multiple standards or indicators in each observation. The observation ratings were aggregated across standards and indicators and across observation cycles to obtain the overall scores or domain scores, although the procedures used to calculate aggregated scores varied by district.

AIR conducted initial analyses of the observation data obtained from each participating district, focusing on distribution of teacher observation scores across teachers and the correlations across standards and observations. The analyses indicated that the distributions of scores for the full teacher sample in each district were adequate for outcome measures and that the correlations across standards and observation cycles suggest reliability appropriate for the evaluation. Table B1, for example, presents some of the statistical properties of the 2014–15 observation scores in each district site. Teachers' scores showed considerable variation in each district, with standard deviations ranging from 0.45 to 0.63. The bivariate correlations between scores for different indicators or standards ranged between 0.50 and 0.79. When aggregated across observations, the reliabilities of scores across the standards or indicators were generally high ranging from 0.88 to 0.95.

The remainder of this appendix describes the teacher observation rubrics used by the four districts and explains how the teacher observation scores obtained from each district were used to construct the outcome measures for the analysis.

Table B1. Overall Classroom Observation Scores in the Evaluation Sites, 2014–15

	Mean[a]	Standard Deviation[a]	Correlations between scores for indicators or standards included		Reliability (Cronbach's Alpha)
			Minimum	Maximum	
Chicago	3.11	0.45	0.61	0.79	0.95
District of Columbia	3.17	0.46	0.61	0.78	0.95
Nashville	3.71	0.63	0.50	0.75	0.95
New Orleans	3.03	0.51	0.54	0.66	0.88

[a] Nashville used a 5-point scale; the other three sites used a 4-point scale.

Observation System Used in Chicago: The Chicago Public Schools Framework for Teaching

The Chicago Public Schools Framework for Teaching evaluated teachers on four domains: (a) Planning and Preparation, (b) Classroom Environment, (c) Instruction, and (d) Professional Responsibilities. Within each domain, teachers were evaluated on several components. Classroom observations focused on the Classroom Environment and Instruction domains (Table B2). Teachers received various numbers of formal and informal observations.[25] During each observation, teachers were rated on a scale of 1 to 4 (1 = unsatisfactory, 2 = basic, 3 = proficient, and 4 = distinguished) on each component within the two domains.

We averaged teachers' scores for each component across observations to produce a single score for each component. Within each domain, the component scores were then averaged to produce a domain score that corresponds to the Classroom Environment and Instruction domains of the Danielson Framework. The overall instructional practice score for each teacher was the weighted average of the two domain scores, with the Instruction domain score accounting for approximately 62% and the Classroom Environment domain score accounting for approximately 38% of the overall score. The weights for the two domains were consistent with the percentage weights used by the Chicago Public Schools when calculating the final teacher observation ratings.[26]

Observation System Used in the District of Columbia: IMPACT Teaching and Learning Framework

As part of the District of Columbia Public Schools' IMPACT system, the Teaching and Learning Framework had three domains or sections: Plan, Teach, and Increase Effectiveness. This evaluation focuses on the Teach domain only. Teachers were evaluated on the Teach domain through formal classroom observations.[27] The Teach domain rubric assessed teachers' performance on nine standards (Table B2). In 2011–12, teachers normally had five formal observations during the course of the year: three by an administrator (principal or assistant principal) and two by third-party observers who were called master educators. In 2012–13 through 2013–14, however, the number of observations teachers received varied from one to six, based on level of teacher experience. Most teachers in the district received four observations.[28] For each formal observation, teachers received a 1 to 4 rating for each standard of the Teach domain.

---

[25] Data received from the Chicago site showed that the number of total observations a teacher received ranged from one to seven in 2012–13 and from one to six in both 2013–14 and 2014–15. If a teacher received more than four observations, then scores from only the top four formal observations or top three formal and one informal observation were used to calculate observation scores (and, hence, only scores from those four observations were provided to AIR).

[26] The Chicago Public Schools used the following percentage weights for each domain when calculating teachers' final observation ratings: 25% for Planning and Preparation, 40% for Instruction, 25% for Classroom Environment, and 10% for Professional Responsibilities. The ratio of the Instruction domain weight to the Classroom Environment domain weight was 40:25, which is the weight ratio that AIR used when combining the two domain scores to obtain the overall instructional practice score.

[27] Teachers also received informal observations, but they were not rated on those observations.

[28] Ninety-four percent of the teachers in the final analytic sample for outcomes in the second year of teaching and 97% of teachers in the final analytical sample for outcomes in the first year of teaching received four or more formal observations.

We averaged teachers' scores for each standard across observations to produce a single score for each standard. The overall instructional practice score for each teacher was the average of the teacher's scores on all nine standards. To construct a domain score that corresponded to the Instruction domain of the Danielson Framework, we calculated the mean of teachers' average scores for standards 1 through 8. To construct a score that corresponded to the Classroom Environment domain of the Danielson Framework, AIR used the average score for standard 9.

Observation System Used by Schools in Tennessee: Tennessee Educator Acceleration Model Rubric

According to Tennessee's state model for teacher evaluation, principals, assistant principals, or instructional leaders used the Tennessee Educator Acceleration Model (TEAM) rubric to observe teachers. The TEAM rubric has four domains: Instruction, Planning, Environment, and Professionalism. The number of observations each teacher received varied based on the teacher's type of teaching license and his or her evaluation results from the prior year.[29] Classroom observations focused on three domains: Planning, Environment, and Instruction. For each observation, teachers received a 1 to 5 rating for each indicator within a domain.

The teacher observation data that AIR received included teachers' ratings for 16 indicators, 4 under the Environment domain and 12 under the Instruction domain. We averaged teachers' scores for the four indicators in the TEAM rubric Environment domain across observations to produce a domain score that corresponded to the Classroom Environment domain of the Danielson Framework; likewise, teachers' scores for the 12 indicators in the TEAM rubric's Instruction domain were averaged across observations to obtain a domain score that corresponded to the Instruction domain of the Danielson Framework. The overall instructional practice scores for teachers were obtained by averaging their scores for all 16 indicators in the TEAM rubric Instruction and Environment domains across all observations.[30]

Although teachers were rated on the TEAM rubric Planning domain during classroom observations, the rating was based on the evaluator's review of teachers' instructional plans, assessment plans, and student work assignments. The rating for this domain did not reflect what was actually happening in classroom instruction; thus, scores from the Planning domain were excluded from our analysis.

[29] The teacher observation data provided to AIR did not include information about the number of observations each teacher received.

[30] This aggregation method is equivalent to weighting the Instruction domain three times (12:4) more than the Environment domain. Instruction accounts for 75% of the overall instructional practice score, and Environment accounts for the remaining 25%.

Observation System Used by Schools in Louisiana: Compass Teacher Rubric

According to the teacher evaluation system in Louisiana, principals, assistant principals, or other trained designees used the state's Compass Teacher Rubric[31] to observe teachers. The Compass Teacher Rubric used five of the 22 components in the Danielson Framework, including one from the Planning and Preparation domain, one from the Classroom Environment domain, and three from the Instruction domain (Table B2). Teachers earned a score of 1, 2, 3, or 4 on each component during each observation. A teacher's score for each observation was the average of his or her scores on the five components of the observation rubric. A teacher's overall professional practice score was the average of the teacher's scores across observations.[32] Compass required a minimum of two observations per year. However, local education agencies (LEAs) determined the number of observations beyond the required two.

Although the Louisiana Department of Education (LDOE) recommended the Compass rubric, some LEAs and charter schools received approval from LDOE to use an alternative tool or rubric.[33] LEAs using an alternative observation tool could use rubrics that were not on a 4-point scale, but they had to ensure that the scores ultimately could be converted to a 4-point scale so that scoring was consistent across LEAs.

The observation data AIR received from LDOE for the 2012–13 school year included only teachers' overall professional practice scores. The department did not provide component or domain scores. Information on the observation rubrics used and the number of observations each teacher received also was not available. However, the data for the 2013–14 and 2014–15 school years included overall professional practice scores, the score for each component in each cycle, information on the rubrics used for each observation, and information on how the scores were aggregated to obtain overall professional scores. Because the more detailed observation data were missing for some Fellows, we decided to use teachers' overall professional practice scores as the only outcome measure in the analyses of teachers' instructional practice across cohorts and years in the New Orleans site.[34]

---

[31] The observation rubric used in this study is one part of Louisiana's comprehensive evaluation system, the Compass Teacher Evaluation System, which includes student growth using student learning targets and value-added measures when available.

[32] Conversations with LDOE staff members suggested that, in some cases, the overall professional practice score was generated based on (but not calculated from) the individual observation ratings. For example, an evaluator may choose to give an overall professional practice rating to a teacher that is based on the evaluator's overall impression during observations rather than averaging scores across observations.

[33] For example, according to e-mail communications with TNTP and LDOE in January 2013, 11 charter schools that had Fellows (2012 cohort) placed during the 2012–13 school year used an alternative rubric approved by the department. These schools employed approximately 25% of the teachers in the 2012 cohort.

[34] The distribution of the overall professional scores followed an approximately normal distribution. In 2013–14, the correlation between the overall professional practice scores and the overall observation scores that we calculated by averaging component scores was .95. We determined that the overall professional practice score showed satisfactory statistical properties to use as an outcome measure.

Table B2. Classroom Observation Domains and Components Used to Measure Instructional Practice

Baltimore: Instructional Framework	Chicago: Framework for Teaching	District of Columbia: Teaching and Learning Framework	Nashville: TEAM Rubric	New Orleans: Compass Teacher Rubric
Domain: Teach 1. Communicate standards-based lesson objectives. 2. Present content clearly. 3. Use strategies and tasks to engage all students in rigorous work. 4. Use evidence-dependent questioning. 5. Check for understanding and provide specific, academic feedback. 6. Facilitate student-to-student interaction and academics. 7. Implement routines to maximize instructional time. 8. Build a positive, learning-focused classroom culture. 9. Reinforce positive behavior, redirect off-task behavior, and de-escalate challenging behavior.	**Domain: The Classroom Environment** 1. Create an environment of respect and rapport. 2. Establish a culture for learning. 3. Manage classroom procedures. 4. Manage student behavior. **Domain: Instruction** 1. Communicate with students. 2. Use questioning and discussion techniques. 3. Engage students in learning. 4. Use assessment in instruction. 5. Demonstrate flexibility and responsiveness.	**Domain: Teach** 1. Lead well-organized, objective-driven lessons. 2. Explain content clearly. 3. Engage students at all learning levels in rigorous work. 4. Provide students with multiple ways to move toward mastery. 5. Check for student understanding. 6. Respond to student understanding. 7. Develop higher-level understanding through effective questioning. 8. Maximize instructional time. 9. Build a supportive, learning-focused classroom community.	**Domain: Environment** 1. Expectations 2. Manage student behavior. 3. Environment 4. Respectful culture **Domain: Instruction** 1. Standards and objectives 2. Motivate students. 3. Present instructional content. 4. Lesson structure and pacing 5. Activities and materials 6. Questioning 7. Academic feedback 8. Group students. 9. Teacher content knowledge 10. Teacher knowledge of students 11. Thinking 12. Problem solving	**Domain: Planning and Preparation** 1. Set instructional outcomes. **Domain: Classroom Environment** 1. Set instructional outcomes. **Domain: Instruction** 1. Use questioning and discussion techniques. 2. Engage students in learning. 3. Use assessment in instruction.

Appendix C. Data Collection

Data Collection

Data were requested on an annual basis from existing district and state data systems for the 2010–11 through 2014–15 school years, depending on which districts were included in the cohorts studied. For the analysis of student achievement, we requested data on students and teachers in tested subjects and tested grades with relevant prior-year test scores. The data requested included student test scores, student demographics, teacher demographics, grades and subjects taught, class rosters, and school-level demographic characteristics and proficiency. For the analysis of teacher instructional practices and teacher retention, we requested classroom observation scores, teacher demographics, grades and subjects taught, and school-level student demographic characteristics for teachers in all subjects and grades within the districts.

We requested data in de-identified formats, with unique random identifiers for students and teachers created by the district or state agency staff prior to the transfer of records to AIR. To comply with data use requirements, in some sites we requested separate files with unique random teacher identifiers for data related to teacher instructional practice, such that teacher data used to examine instructional practice were not directly linkable to teacher data used to examine student achievement.

For the evaluation of student achievement, we obtained data from the one participating site with a 2010 cohort (District of Columbia), all five participating sites with a 2011 cohort (Charlotte, District of Columbia, Fort Worth, Nashville, and New Orleans), data from all seven participating sites with a 2012 cohort, and data from six participating sites with a 2013 cohort. For the analyses of instructional practice and teacher retention, four districts with eligible teacher observation rubrics provided the necessary data.

Each participating district provided AIR with data files uploaded through AIR's secure file transfer protocol site. AIR staff conducted an initial review of the data, checking for missing variables and values and other data issues (e.g., variable and value definitions) and verifying the linkability of students to their teachers. AIR then communicated with each district to resolve any identified issues and request supplemental data if needed.

To accurately flag Fellows in the target cohorts in the district data files in the event of incomplete existing district records on the teacher certification program, while maintaining de-identified records for the evaluation, we used a process in which TNTP provided lists of Fellows for district data personnel. The district personnel identified these individuals in district records and included variables indicating Fellows in the de-identified files provided to AIR.[35] We used a similar process to flag Teach For America (TFA)

[35] Across the sites and cohorts (2011, 2012, 2013) with data available for the evaluation, according to TNTP, 1,642 Fellows entered inservice training. Based on the lists of Fellows provided by TNTP, district staff flagged 1,204 individual teachers as Fellows in district records provided to AIR. The differences between TNTP's records and districts' records could be caused by Fellows not completing the employment process in these districts or not being accurately identified in district records provided to AIR. We cannot verify why some individuals from TNTP's records were unflagged in the district files, but communication with district staff indicated that a failure to flag a true Fellow from TNTP's records would be an infrequent occurrence. It is possible that some Fellows who successfully entered district employment were not accurately flagged in the data provided to AIR and remained in a pool of 17,487 potential comparison teachers.

corps members in district files from records provided by TFA. Corps members participated in TNTP inservice training at some sites. As such, it was necessary to flag and exclude corps members from the comparison pool as much as possible based on TFA's records of recent cohorts.

Data Preparation

For each year of analysis, AIR followed the same process to prepare the data for matching and analysis, including cleaning the data (e.g., removing duplicate records), recoding and renaming variables for consistency across districts, and merging files within and across districts. For the student achievement analysis, we eliminated from the universe of teachers in each district those who taught nontested grades and subjects, prior cohorts of TNTP teachers and TFA corps members, and comparison teachers in grades and subjects that do not have Fellows counterparts. For the analysis of teacher instructional practice outcomes, the following teachers were removed from the potential comparison pool: teachers without classroom observation scores, TFA teachers, teachers from other Teaching Fellows cohorts (other than the analysis cohort), teachers with no data on years of experience, and teachers who taught grade levels and subjects that did not have Fellows counterparts. One additional step was applied in refining the New Orleans sample. The New Orleans site included multiple LEAs, including charter management organizations that implemented the Teaching Fellows program. Because there was no single district to define for the evaluation, the sample was restricted to teachers at schools that hired at least one Fellow from any of the three cohorts. The matched sample of teachers and students was constructed from the remaining pool of teachers.

Appendix D. Analytic Methods

Selection Models for the Matching Process

Because Fellows were newer to the teaching profession than the larger population of teachers, teaching experience was a key characteristic that needed to be balanced between the two groups. For the analyses of both instructional practice and student achievement, we used experience as a conditioning variable to create potential comparison teachers who had the same level of prior experience as the Fellows. This method imposed sample inclusion restrictions on comparison teachers on the basis of their experience, creating two different pools of potential comparison teachers: (a) a pool of comparison teachers with one prior year of experience for the sample of Fellows in their second year of teaching and (b) a pool of comparison teachers with zero prior years of experience for the sample of first-year Fellows. Because within each selection pool, Fellows and comparison teachers had the same level of experience, it was not necessary to control for experience in the selection models for matching teachers and classes.

Teacher Instructional Practice

To construct the analytic samples for instructional practice, matching for the second-year teachers was conducted independently of the matching for the first-year teachers, with matching implemented within cohort and district for each sample. We estimated propensity scores using a logistic regression model, in which the outcome variable was an indicator of whether a teacher was trained through the Teaching Fellows program, and the predictors were teacher characteristics and the characteristics of schools where the teachers taught. The propensity score for a teacher represents the probability that a teacher was trained through the Teaching Fellows program, given the observed characteristics. Then each Fellow was matched to at most two comparison teachers in the same grade level with the closest propensity scores (within a caliper of 1 *SD*)—that is, the two nearest "neighbors."

The general form of the propensity score (logistic regression) model used for matching teachers was as follows:

$$logit(P(Fellow)) = \eta + \sum \lambda Z + \sum \theta C$$

where *Fellow* is an indicator of whether a teacher was trained through the Teaching Fellows program (*Fellow* = 1 if teacher is a Fellow or 0 otherwise), *P*(*Fellow*) is the propensity of a teacher to be trained through the Teaching Fellows program, η is an intercept, Z is a set of teacher background characteristics (i.e., age, racial minority status, gender, grade, subject, and missing data indicators when applicable), C is a set of school background characteristics (i.e., the percentage of students who are English language learners, the percentage of students who are eligible for free or reduced-price lunch, the percentage of students in special education programs, the percentage of students proficient in reading, and the percentage of students proficient in mathematics), λ is a set of coefficients that represents the association between each teacher characteristic and the logit of the propensity score, and θ is a set of coefficients that represents the association between each school characteristic and the logit of the propensity score.

Student Achievement

To construct the analytic samples for student achievement, matching for the second-year teachers was conducted independently of the matching for the first-year teachers. For each sample, matching was implemented in three steps, within cohort and district, as follows.

First, we estimated propensity scores by using a logistic regression model, in which the outcome variable was an indicator of whether a class is taught by a Fellow, and the predictors were baseline teacher and classroom characteristics. The propensity score for each class represented the probability that a class was taught by a Fellow, given the observed characteristics. We matched each class taught by a Fellow to at most two classes in the same grade3[36] and content area taught by comparison teachers with the closest propensity scores (within a caliper of 1 *SD*)—that is, the two nearest neighbors.

Second, we estimated propensity scores for students by using a logistic regression model with an indicator of whether a student is taught by a Fellow as the outcome and observed student characteristics as the covariates. Within each group of matched teachers, each student taught by a Fellow was matched to at most two students in the same grade and attending similar classes[37] but taught by a comparison teacher with the closest propensity scores (within a caliper of 1 *SD*). After matching separately by grade and subject within each cohort and district, we combined the matches across grades, subjects, cohorts, and districts.

The general form of the propensity score (logistic regression) model used for matching classes was as follows[38]:

$$logit(P(Fellow)) = \eta + \sum \lambda Z + \sum \theta C + \varphi R$$

where *Fellow* is an indicator of whether a class was taught by a Fellow (*Fellow* = 1 if teacher is a Fellow or 0 if not); *P*(*Fellow*) is the propensity of a class to be taught by a Fellow; η is an intercept; Z is a set of baseline teacher characteristics (age, racial minority status, and gender); C is a set of baseline classroom characteristics (average of students' prior-year test scores [*z*-scores] in the same subject, grade level, subject, the percentage of students who are ELLs, the percentage of students who belong to racial or ethnic minority groups, the percentage of students who are eligible for free or reduced-price lunch, the percentage of students who have an IEP, and the percentage of students who are female); R is the school percentage proficiency rate in a given grade and subject[39]; λ is a set of coefficients that represents the association between each baseline teacher characteristic and the logit of the

---

[36] To simplify the process, if a class (e.g., in high school) contained students in different grades, the average grade of the students in the class (rounded to a whole number) was used.

[37] In elementary and middle schools (Grades 4–8), students of Fellows were compared with similar students in the same grade. In high school, because students may take the same course while in different grades, comparison students were not necessarily in the same grade as students of Fellows. In middle and high schools, the matching classes were in the same content area (e.g., algebra and middle school mathematics) but were not necessarily the same courses.

[38] The propensity score model for matching classes also included missing data indicators for covariates with missing values.

[39] School percentage proficiency rates were not provided for some cohorts and districts. Because it was not available for all cohorts and districts, it was not used in the analytic models for both first-year teachers and second-year teachers.

propensity score[40]; θ is a set of coefficients that represents the association between each baseline classroom characteristic and the logit of the propensity score; and φ is a coefficient that represents the association between school percentage proficiency rate and the logit of the propensity score.

For matching students within classes, we used the following specification for the propensity score model[41]:

$$logit(P(Fellow)) = \pi + \sum \beta S$$

where *Fellow* is an indicator of whether a student was taught by a Fellow; P(*Fellow*) is the propensity of a student to be taught by a Fellow; π is an intercept; S is a set of baseline student characteristics (prior-year test score [*z*-score], grade, subject, ELL status, IEP indicator, free or reduced-price lunch eligibility, male indicator, and racial minority status); and β is a set of coefficients that represents the association between each baseline student characteristic and the logit of the propensity score.

Last, we assessed whether the matching process produced groups that were statistically equivalent in baseline characteristics. Consistent with the WWC standards (WWC, 2013), we considered the two groups to be balanced if the standardized average difference in prior student achievement between the two groups was less than or equal to 0.25 *SD* in the combined sample. However, we also checked the balance on student and teacher demographic characteristics. These checks of baseline equivalence were conducted on the full sample of first-year teachers and the full sample of second-year teachers as well as separately by district within each sample. These checks indicated that matching produced equivalent groups, particularly in prior achievement.

A few things should be noted regarding the propensity score models. First, ideally when matching teachers and classes, separate selection models are used for each subject and grade. However, this was not feasible given the small numbers of Fellows in each subject and grade. Instead, a single propensity score model was used to generate propensity scores by combining classes across grades and subjects in each district. The grade and subject of the class were accounted for by including them as predictors in the selection model. In addition, after the propensity scores were generated, classes were matched within the same grade and subject.

Second, in the matching of students, the goal was to use the full set of available student covariates from each district (e.g., baseline achievement score [z-score] in mathematics, reading, science, or social studies; ELL indicator; minority indicator; free or reduced-price lunch eligibility indicator; IEP indicator; female indicator; and age). However, the actual covariates used in estimating propensity scores varied within each group of matched teachers and classes. In some cases, covariates were excluded from the selection model either because of perfect collinearity (e.g., students who were eligible for free or reduced-price lunch also were minorities) or because they predicted the outcome perfectly (e.g., all female students happened to be taught by a Fellow, and all male students happened to be taught by a comparison teacher). Also, because a separate selection model was run for each group of matched classes, in some cases, the number of students within each group of matched classes was too small for us to obtain stable propensity score estimates. In these cases, the number of covariates used was limited to the pretest score only.

---

[40] The logit of the propensity score is equal to the log of the following ratio: propensity score divided by 1 minus the propensity score.

[41] The propensity score model for matching students also included missing data indicators for covariates with missing values.

Third, matching of students was implemented with replacement. This means a student of a comparison teacher can potentially be matched to more than one student of a Fellow. This method produces matches that have more similar propensity scores because students of Fellows can be matched to the nearest comparison student even if that comparison student already has been matched previously. In the analysis phase, these comparison students were given a weight of 1 (as all other students were), no matter how many times they were matched.

Last, to achieve better balance, the common support restriction was imposed during the propensity score matching for students—that is, treatment cases with propensity scores higher than the maximum or lower than the minimum propensity score of comparison students were dropped from the selection pool.

Models for Estimating Group Differences

This section details the analytic strategies and statistical models used to estimate two differences: (a) differences in instructional practice between Fellows and matched comparison teachers and (b) differences in achievement between students of Fellows and students of matched comparison teachers. As previously stated, the analysis for instructional practice was conducted independently of the analysis for student achievement. Therefore, the set of teachers included in the analytic samples for student achievement may intersect, but it is not necessarily the same as the set of teachers included in the analytic samples for student achievement.[42]

The analysis for instructional practice was conducted separately for each district, whereas the analysis for student achievement was conducted on the pooled sample across districts. The overall effect on instructional practice was obtained by taking a weighted average of estimates from within-district analyses. In contrast, the overall impact on student achievement was obtained by using district-specific estimates from the combined sample across all districts and then taking a weighted average of the district-specific impacts.

Teacher Instructional Practice

Model for Within-District Analyses

Within each district, we conducted parallel analyses of the first-year and second-year teacher samples using teacher instructional practice scores as outcomes and teacher and school characteristics as covariates. We used an analysis of covariance model to assess whether significant differences existed between the Fellows and the comparison teachers on the outcomes. The analysis of covariance model took the following basic form:

$$Y_k = \pi + \sum \emptyset X_k + \sum \delta Z_k + \beta FELLOW_k + \omega Cohort_{\mathrm{k}} + e_k$$

where Y_k is the score on an instructional practice outcome measure for teacher k; is the intercept or average score on the outcome measure for comparison teachers (in the reference cohort if the

[42] For example, the analytic samples for student achievement included only tested grades and subjects, whereas the analytic samples for instructional practice included nontested grades and subjects.

analysis includes multiple cohorts); X_k is a set of teacher characteristics to serve as covariates, including teacher age, grade, subject area, gender, minority status, and missing data indicators (if applicable); Z_k is a set of school characteristics, including the percentage of students eligible for free or reduced-price lunch, the percentage of ELL students, the percentage of students in special education programs, the percentage of students from minority backgrounds, school enrollment, proficiency rates on standardized reading and mathematics assessments, and missing data indicators (if applicable); $Fellow_k$ is a dummy indicator for whether teacher *k* received training through the Teaching Fellows program; $Cohort_k$ is a dummy variable(s) for cohort (when the sample included teachers from more than one cohort); and e_k is a random error term associated with teacher *k*, assumed to be independently and identically distributed.

Given this equation, the estimate for the effect of the Teaching Fellows program on a given teacher outcome is represented by β; ω represents the differences between cohorts on the outcome, and $\emptyset$ and δ denote regression coefficients for the teacher and school characteristics included in the model.

From the regression model for each outcome, we calculated the adjusted means for the Fellows and the comparison teachers. The difference in the adjusted means on an outcome measure was converted into an effect-size estimate that represents the difference in outcomes scores in *SD* units (Hedges' *g*). The computation of Hedges' *g* was derived from equations specified in version 3.0 of the WWC Procedures and Standards Handbook (WWC, 2013) for covariate-adjusted regression models:

$$g = \frac{\omega(x_i^{'} - x_c^{'})}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

where $x_i^{'}$ is the covariate-adjusted mean on the outcome for the Fellows; $x_c^{'}$ is the covariate-adjusted mean on the outcome for the comparison teachers; ω is Hedges' correction factor for small samples, where $[1 - 3/(4N - 9)]$ and *N* is total sample size; n_1 is the number of Fellows; n_2 is the number of teachers trained through other types of programs; s_1^2 is the variation in outcome measure for Fellows; and s_2^2 is the variation in outcome measure for teachers trained through other types of programs. Per the *WWC Procedures and Standards Handbook* (WWC, 2013), this calculation of effect size used the correct group difference estimate (i.e., the numerator) and the correct estimates for within-teacher variance.

Summarizing Results Across Districts

To assess the overall impact of the Teaching Fellows program on a given measure of teacher instructional practice across districts, we computed a summary effect that was essentially a weighted average of district-specific effect sizes. We followed the methods for performing the fixed-effect meta-analysis described by Borenstein, Hedges, Higgins, and Rothstein (2009). The weight assigned to each district estimate is the inverse of the district estimate's variance (i.e., precision weight):

$$W_i = 1/V_{Gi}$$

where V_{Gi} is the variance for the effect size estimate G_i in district *i*.

The weighted average was then computed as follows:

$$\text{Weighted Average } (T) = \frac{\sum W_i G_i}{\sum W_i}$$

The variance of the summary effect T was estimated as the reciprocal of the sum of the weights ($V_T = 1/\sum W_i$),and the standard error of the summary effect (SE_T) was the square root of the variance (i.e., $\sqrt{V_T}$). The 95% confidence interval for the summary effect could then be computed as follows:

$$\text{Lower Limit} = T - 1.96 \times SE_T$$

$$\text{Upper Limit} = T + 1.96 \times SE_T$$

A Z value to test the null hypotheses that the true common effect was zero could be computed as follows:

$$Z = \frac{T}{SE_T}$$

For a two-tailed test, the p value was given by

$$p = 2[1 - (\Phi(|Z|)]$$

where $\Phi(|Z|)$ is the standard normal cumulative distribution.

Student Achievement

This section provides details of the analytic strategies used to estimate the differences in achievement between students of Fellows and comparable students of matched comparison teachers. It describes the analytic samples, the analyses conducted, the outcomes and covariates used, the statistical models used to estimate group differences, and the estimation of overall group differences. We conducted parallel analyses of the first-year and second-year samples of teachers and students.

Outcomes

We used as outcomes students' scores in 2011–12, 2012–13, 2013–14, or 2014–15 state assessments in tested subjects (mathematics, reading, science, or social studies) and tested grades (Grades 4–12). Because students' scores were combined across subjects, grades, districts, and cohorts, test scores were converted to the same scale (z-score) by standardizing scores separately by cohort, district, subject, and grade. A z-score was calculated as the difference between a student's raw score and the district average raw score for a particular subject, grade, and cohort, divided by the district *SD* of raw scores for that subject, grade, and cohort. Hence, a student's z-score could be interpreted as the student's performance relative to the district average in that particular grade, subject, and cohort.

Covariates

We used covariates to account for baseline differences at the student, classroom, and teacher levels. At level 1, we used students' prior year score (z-score) in state assessments, grade, subject, ELL status, minority status, an IEP status, and gender. At level 2, we incorporated variables that described

classroom context: mean pretest scores, the percentage of students who are ELL, the proportion who are minority, the proportion who have an IEP, and the proportion who are female. At level 3, we accounted for gender and race (indicator for White race). District indicators (for each of the seven districts) also were added at level 3 to adjust for (observed and unobserved) differences across districts. To remove any residual differences in the background characteristics, we also included at level 1 the logit of a student's estimated propensity score.

Statistical Model

The differences between Fellows and comparison samples were estimated using three-level hierarchical linear models with students (level 1) nested within classrooms (level 2) nested within teachers (level 3). These models accounted for similarities among students who belonged to the same classroom and similarities among classes taught by the same teacher. District-specific group differences from the full sample were pooled into an overall group difference as discussed later. The general form of the three-level analytic model is as follows.

Level 1: Student Level

$$Y_{ijk} = \pi_{jk} + \sum \beta S_{ijk} + e_{ijk}$$

where Y_{ijk} is the achievement score (z-score) for student *i* in class *j* with teacher k; S_{ijk} is a set of baseline student characteristics (prior-year test score [z-score], ELL status, IEP indicator, female indicator, and racial minority status) and student's grade and subject; π_{jk} is the student-level intercept, or average achievement of students in class *j* for teacher *k*; β is a set of coefficients that represents the association between student achievement and each baseline student characteristic, grade, and subject; and e_{ijk} is student-level residual error, assumed to be independently and identically distributed.

Level 2: Classroom Level

$$\pi_{jk} = \gamma_k + \sum \alpha C_{jk} + u_{jk}$$

where γ_k is the average achievement of students across all classes taught by teacher *k*; C_{jk} is a set of baseline characteristics for class *j* taught by teacher *k* (average of students' prior-year test scores (*z*-scores) in the same content area, the percentage of students who are ELLs, the percentage of students who belong to racial or ethnic minority groups, the percentage of students who have an IEP, and the percentage of students who are female); α is a set of coefficients that represents the association between average class achievement and each baseline classroom characteristic; and u_{jk} is classroom-level residual error, assumed to be independently and normally distributed.

Level 3: Teacher Level

$$\gamma_k = \sum \eta_m D_m + \sum \theta_m D_m Fellow_k + \sum \lambda Z_k + v_k$$

where D_m is an indicator for district *m* (*m* = 1 to *M*, corresponding to the *M* study districts); $Fellow_k$ is an indicator of whether teacher *k* is a Fellow ($Fellow_k$ = 1 if teacher *k* is a Fellow or 0 if not); Z_k is

a set of baseline characteristics (teacher's racial minority status and gender) and a cohort indicator for teacher *k*; η_m is the overall mean achievement of students across all classes and teachers in district *m*; θ_m is the difference in the average achievement between the two groups of students in district *m*; λ is a set of coefficients that represents the association between student achievement and teacher characteristics; and v_k is teacher-level random error assumed to be independent and normally distributed.

In the model here, the key parameter is θ_m, which represents the average program-related difference in district *m*; η_m is the regression-adjusted mean achievement for students taught by comparison teachers in district *m*, and $\eta_m + \theta_m$ is the regression-adjusted mean achievement of students taught by Fellows in district *m.*

Estimating Overall Group Differences

In each analytic sample, the estimated overall group difference was a weighted average of the estimated district-specific group differences $\{\theta_m\}$ and can be interpreted as the average difference between students of Fellows and those of comparison teachers in our study districts. Because the Teaching Fellows program was targeted at the teacher level, we weighted district-specific group differences by the proportion of Fellows per district in the analytic samples. Specifically, an estimate of the overall impact θ was calculated as $\hat{\theta} = \sum_{m=1}^{M} w_m \hat{\theta}_m$, where the weight $w_m = n_m / \left(\sum_{k=1}^{M} n_m\right)$ and n_m is the number of Fellows from each district. Similarly, the weighted mean of $\hat{\eta}_m$ represented the estimated overall mean achievement of students taught by comparison teachers, and the weighted mean of ($\hat{\eta}_m + \hat{\theta}_m$) was the estimated overall mean achievement of students taught by Fellows.[43]

Estimating district-specific group differences instead of assuming a constant group difference across districts has the following advantages: (a) It accounts for differences in policies, context, and characteristics, as well as differences in the implementation of the program across districts; (b) it conforms to the matching design used—namely, matching teachers and students separately by district; (c) it takes advantage of the blocked matching design by comparing students taught by Fellows with only matched comparison students within the same district; and (d) it accounts for differences in the precision of site-specific estimates that result from sample size variations across districts.

To summarize

- The estimates come from a three-level hierarchical linear model.
- The estimates are adjusted for all the covariates specified previously.
- The overall group differences are averages of district-specific group differences weighted by the proportion of Fellows from each district in the analytic sample.

[43] A similar weighting procedure was applied in computing the standard error of the estimates. For example, the standard error for the overall estimate of group difference θ was given by $\sqrt{\left(\sum w_m^2 Var(\hat{\theta}_m)\right)}$, where $Var(\hat{\theta}_m)$ is the variance of the estimated group difference in the district *m*.

Methods for Comparing Teacher Retention

The district data were used to identify Fellows who began teaching in 2011, 2012, and 2013 plus teachers from other training programs who were hired by the districts during the year that Fellows entered inservice training. F44 Teachers who had been flagged in district records as TFA corps members or former corps members were excluded from the sample. Table E4 provides sample sizes, overall and by cohort.

A longitudinal file was constructed for each cohort to track individual teachers from when they began teaching through academic year 2014–15. Dichotomous indicators were created to flag whether teachers returned for the second, third, and fourth year to a teaching position within the same district. The determination of teachers' retention was based on whether they were present in the district teacher records in each subsequent year.

To calculate site-specific retention rates, the number of teachers within each group (Fellows or other teachers) in each site who returned for the nth year and all previous years was divided by the number of teachers in the beginning. The site-specific retention rates then were pooled into a weighted average, using the proportion of Fellows in each site as the weight, to generate estimates of the overall teacher retention rates for Fellows and other teachers. Pearson's chi-squared tests were calculated to determine whether the observed differences were statistically significant. The statistical power to detect differences declined across time because the sample size decreased as members of the original cohorts left their positions.

[44] The New Orleans site included multiple LEAs, including charter management organizations, that implemented the Teaching Fellows program. Because there was no single district to define for the evaluation, the sample was restricted to teachers at schools that hired at least one Fellow from any of the three cohorts.

Appendix E. Evaluation Samples

This appendix provides details about the evaluation samples used in the teacher outcomes and student outcomes analyses. Tables E1–E4 provide sample sizes, overall and by cohort. Tables E5–E8 present baseline characteristics of each analysis sample.

Table E1. Number of Students, Classes, and Teachers in the Analytic Sample for Student Achievement for Teachers in Their First Year of Teaching, by District

District	Overall			2011 Cohort			2012 Cohort			2013 Cohort		
	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total
Overall												
Students	18,826	18,273	37,099	4,568	3,917	8,485	6,159	5,274	11,433	8,099	9,082	17,181
Classes	840	1,393	2,233	197	312	509	294	490	784	349	591	940
Teachers	445	1,014	1,459	118	245	363	152	352	504	175	417	592
Baltimore												
Students	287	239	526				287	239	526	X	X	X
Classes	14	22	36				14	22	36	X	X	X
Teachers	8	14	22				8	14	22	X	X	X
Charlotte												
Students	3,396	2,770	6,166	822	509	1,331	375	440	815	2,199	1,821	4,020
Classes	99	172	271	25	43	68	25	47	72	49	82	131
Teachers	67	145	212	18	39	57	10	36	46	39	70	109
Chicago												
Students	1,999	2,217	4,216				1,069	1,151	2,220	930	1,066	1,996
Classes	177	300	477				96	174	270	81	126	207
Teachers	65	211	276				36	127	163	29	84	113

District	Overall			2011 Cohort			2012 Cohort			2013 Cohort		
	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total
District of Columbia												
Students	857	702	1,559	204	182	386	202	180	382	451	340	791
Classes	59	87	146	23	33	56	20	29	49	16	25	41
Teachers	32	64	96	11	26	37	11	20	31	10	18	28
Fort Worth												
Students	4,722	5,044	9,766	1,330	1,016	2,346	1,461	1,064	2,525	1,931	2,964	4,895
Classes	207	297	504	59	81	140	61	79	140	87	137	224
Teachers	103	168	271	30	50	80	35	38	73	38	80	118
Nashville												
Students	4,162	3,919	8,081	1,640	1,516	3,156	1,520	1,325	2,845	1,002	1,078	2,080
Classes	150	272	422	56	96	152	48	92	140	46	84	130
Teachers	83	213	296	37	76	113	28	74	102	18	63	81
New Orleans												
Students	3,403	3,382	6,785	572	694	1,266	1,245	875	2,120	1,586	1,813	3,399
Classes	134	243	377	34	59	93	30	47	77	70	137	207
Teachers	87	199	286	22	54	76	24	43	67	41	102	143

Note. Gray shading represents a site that did not implement the Teaching Fellows program for this cohort and school year. X represents a site that implemented the Teaching Fellows program, but the data necessary to analyze student achievement were not available. Data on the first year of teaching of the 2010 cohort from the District of Columbia Public Schools were not included in this study. The list of 2010 cohort Fellows provided by the district could not be verified with the program records provided by TNTP in 2011–12 because of inconsistencies in teacher identifiers in the 2010–11 and 2011–12 district files.

Source. AIR's analysis is based on data provided by districts.

Table E2. Number of Students, Classes, and Teachers in the Analytic Sample for Student Achievement for Teachers in Their Second Year of Teaching, by District

	Overall			2010 Cohort			2011 Cohort			2012 Cohort			2013 Cohort		
District	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total
Overall															
Students	12,795	10,778	23,573	54	49	103	3,401	2,621	6,022	5,259	4,815	10,074	4,081	3,293	7,374
Classes	587	970	1,557	5	6	11	146	265	411	720	508	1,228	171	281	452
Teachers	303	693	996	2	3	5	69	152	221	137	330	467	95	208	303
Baltimore															
Students	575	554	1,129							575	554	1,129	X	X	X
Classes	49	83	132							49	83	132	X	X	X
Teachers	25	60	85							25	60	85	X	X	X
Charlotte															
Students	2,531	1,893	4,424				53	53	106	810	659	1,469	1,668	1,181	2,849
Classes	79	130	209				6	8	14	23	46	69	50	76	126
Teachers	53	105	158				3	6	9	14	39	53	36	60	96
Chicago															
Students	1,406	1,355	2,761							674	629	1,303	732	726	1,458
Classes	128	229	357							63	117	180	65	112	177
Teachers	48	163	211							23	83	106	25	80	105

	Overall			2010 Cohort			2011 Cohort			2012 Cohort			2013 Cohort		
District	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total	Fellows	Comparison Teachers	Total
District of Columbia															
Students	630	554	1,184	54	49	103	304	284	588	168	140	308	104	81	185
Classes	35	57	92	5	6	11	18	31	49	9	14	23	3	6	9
Teachers	23	46	69	2	3	5	11	24	35	7	13	20	3	6	9
Fort Worth															
Students	3,577	2,718	6,295				1,829	1,375	3,204	714	525	1,239	1,034	818	1,852
Classes	160	243	403				85	132	217	35	47	82	40	64	104
Teachers	73	145	218				30	77	107	21	25	46	22	43	65
Nashville															
Students	2,468	2,204	4,672				698	458	1,156	1,227	1,259	2,486	543	487	1,030
Classes	84	147	231				24	39	63	47	85	132	13	23	36
Teachers	48	107	155				15	29	44	24	59	83	9	19	28
New Orleans															
Students	1,608	1,500	3,108				517	451	968	1,091	1,049	2,140		X	X
Classes	52	81	133				13	18	31	39	63	102	X	X	X
Teachers	33	67	100				10	16	26	23	51	74	X	X	X

Note. Gray shading represents a site that did not implement the Teaching Fellows program for this cohort and school year. X represents a site that implemented the Teaching Fellows program, but the data necessary to analyze student achievement were not available.

Source. AIR's analysis is based on data provided by districts.

Table E3. The Number of Teachers in the Analytic Samples for Teacher Instructional Practice

	Overall			2010 Cohort			2011 Cohort			2012 Cohort			2013 Cohort		
	Fellows	Comparison Group	Total	Fellows	Comparison Group	Total	Fellows	Comparison Group	Total	Fellows	Comparison Group	Total	Fellows	Comparison Group	Total
First-Year Teacher Sample															
Chicago	94	186	280							48	96	144	46	90	136
District of Columbia	214	359	573	X	X	X	96	162	258	76	127	203	42	70	112
Nashville	94	180	274				28	52	80	47	91	138	19	37	56
New Orleans	85	161	246				X	X	X	58	109	167	27	52	79
Total	487	886	1,373	X	X	X	124	214	338	229	423	652	134	249	383
Second-Year Teacher Sample															
Chicago	84	158	242							39		116	45	81	126
District of Columbia	222	368	590	39	69	108	83	136	219	54		132	46	85	131
Nashville	73	130	203				34	52	86	29		87	10	20	30
New Orleans	73	133	206				35	70	105	9		22	29	50	79
Total	452	789	1,241	39	69	108	152	258	410	131		357	130	236	366
Third-Year Teacher Sample															
Chicago	40	78	118							40		118	X	X	X
District of Columbia	137	218	355	30	49	79	62	97	159	46		118	X	X	X
Nashville	44	77	121				22	38	60	22		61	X	X	X
New Orleans	16	31	47				7	14	21	9		26	X	X	X
Total	237	404	641	30	49	79	91	149	240	117		323	X	X	X

Note. Three sites (Baltimore, Charlotte, and Fort Worth) were not included in the analysis of instructional practice outcomes. Gray shading represents a site that did not implement the Teaching Fellows program for this cohort and school year. X represents a site that implemented the Teaching Fellows program, but the data necessary to analyze teacher instructional practice were not available for conducting analysis for the given cohort and school year.

Source. AIR's analysis is based on data provided by districts.

Table E4. The Number of Teachers in the Cohorts Analyzed for Teacher Retention

	Overall			2011 Cohort			2012 Cohort			2013 Cohort		
	Fellows	Other Teachers	Total	Fellows	Other Teachers	Total	Fellows	Other Teachers	Total	Fellows	Other Teachers	Total
Chicago	103	3,029	3,132				57	2,010	2,067	46	1,019	1,065
District of Columbia	253	1,185	1,438	110	385	495	84	325	409	59	475	534
Nashville	106	2,059	2,165	35	606	641	49	657	706	22	796	818
New Orleans	145	1,756	1,901	61	828	889	84	928	1,012	X	X	X
Total	607	8,029	8,636	206	1,819	2,025	274	3,920	4,194	127	2,290	2,417

Note. Gray shading represents a site that did not implement the Teaching Fellows program for this cohort and school year. X represents a site that implemented the Teaching Fellows program, but the data necessary to analyze teacher retention were not available for conducting analysis for the given cohort. This table provides the numbers of teachers in the original cohorts in the first year of teaching.

Source. AIR's analysis is based on data provided by districts.

Table E5. Baseline Characteristics of Students and Teachers Included in the Analytic Samples for Student Achievement

Characteristic	Second-Year Teachers Sample				First-Year Teachers Sample			
	Mean		Mean Difference	Standardized Difference	Mean		Mean Difference	Standardized Difference
	Fellows	Comparison Group			Fellows	Comparison Group		
Student Characteristics								
Baseline achievement (average *z*-score)	-0.23	-0.25	0.02	0.03	-0.27	-0.27	0.00	0.00
ELL	20.6	21.0	-0.3	-0.01	17.6	18.4	-0.8	-0.02
Minority	86.5	85.6	0.9	0.02	87.4	86.9	0.4	0.01
Eligible for free or reduced-price lunch	72.4	72.4	0.1	0.00	74.2	74.3	-0.1	0.00
IEP	16.1	17.3	-1.2	-0.04	11.6	11.3	0.3	0.01
Female	49.7	48.8	1.0	0.02	49.0	49.7	-0.7	-0.01
Teacher Characteristics								
Experience (average in years)	1.0	1.0	0.0	0.00	0.0	0.0	0.0	0.00
Age (average in years)	26.2	27.4	-1.2[a]	-0.17	22.4	24.0	-1.6[a]	-0.22

Note. Means and differences were regression adjusted to account for the clustering of students within classrooms, within teachers, and weighted by the proportion of Fellows from each district in the analytic sample. Because of rounding, the reported difference may differ slightly from the difference between the reported means. The standardized difference was computed by dividing the difference by the pooled *SD* of the characteristic for the two groups (Hedges' *g*). Charlotte did not provide data on eligibility for free or reduced-price lunch, and New Orleans did not provide data on teacher age, so the comparisons for free or reduced-price lunch and teacher age excluded Charlotte and New Orleans, respectively.

[a] Difference is statistically significant at the .05 significance level.

Source. AIR's analysis is based on data provided by districts.

Table E6. Baseline Characteristics of Teachers Included in the Analytic Samples for Teacher Instructional Practice in the First Year of Teaching

Characteristics	Chicago (2012 and 2013 Cohorts Combined)			District of Columbia (2011, 2012, and 2013 Cohorts Combined)			Nashville (2011, 2012, and 2013 Cohorts Combined)			New Orleans (2012 and 2013 Cohorts Combined)		
	Mean			Mean			Mean			Mean		
	Fellows (*n* = 94)	Comparison (*n* = 196)	*g*	Fellows (*n* = 214)	Comparison (*n* = 359)	*g*	Fellows (*n* = 94)	Comparison (*n* = 180)	*g*	Fellows (*n* = 85)	Comparison (*n* = 161)	*g*
Teacher Characteristics												
Age (years)	26.2	26.4	-0.05	26.4	27.1	-0.13	28.0	28.3	-0.05	X	X	X
White	74.5	70.9	0.08	54.0	51.6	0.05	76.7	73.3	0.08	70.6	73.3	-0.06
Female	74.4	73.7	0.02	76.4	77.3	-0.02	71.3	73.3	-0.04	69.4	70.2	-0.02
School Characteristics												
Free or reduced-price lunch eligible	90.9	89.9	0.07	84.8	85.6	-0.04	79.7	80.5	-0.04	85.0	85.3	-0.03
Non-White	95.4	94.1	0.12	95.9	95.8	0.01	72.0	73.1	-0.06	89.5	90.3	-0.06
ELL	30.3	34.0	-0.15	14.1	14.0	0.01	15.7	15.1	0.03	3.6	3.7	-0.01
Special education	14.4	14.5	-0.01	16.3	16.5	-0.02	12.7	12.7	0.01	11.5	11.5	0.01
Reading proficiency	45.1	45.2	0.00	38.7	37.8	0.05	34.6	33.5	0.08	56.6	56.7	0.00
Mathematics proficiency	49.2	49.0	0.01	40.7	40.4	0.02	29.8	29.2	0.05	53.0	53.0	0.00
School size (average enrollment)	901.7	956.6	-0.09	463.5	462.0	0.01	720.7	716.0	0.01	683.5	643.2	0.12

Note. X indicates that data on this variable were not available. *g* is the difference between groups in *SD* units. Differences were adjusted for each cohort. The sample size for some characteristics might be smaller because of missing data. Positive numbers for school characteristics represent the percentages of students in the school.

Source. AIR's analysis is based on data provided by the Chicago Public Schools, the District of Columbia Public Schools, Metropolitan Nashville Public Schools, and the Louisiana Department of Education.

Table E7. Baseline Characteristics of Teachers Included in the Analytic Samples for Teacher Instructional Practice in the Second Year of Teaching

Characteristics	Chicago (2012 and 2013 Cohorts Combined)			District of Columbia (2010, 2011, 2012, and 2013 Cohorts Combined)			Nashville (2011, 2012, and 2013 Cohorts Combined)			New Orleans (2011, 2012, and 2013 Cohorts Combined)		
	Mean			Mean			Mean			Mean		
	Fellows (*n* = 84)	Comparison (*n* = 158)	*g*	Fellows (*n* = 222)	Comparison (*n* = 368)	*g*	Fellows (*n* = 73)	Comparison (*n* = 130)	*g*	Fellows (*n* = 73)	Comparison (*n* = 133)	*g*
Teacher Characteristics												
Age (years)	27.1	27.6	-0.09	27.2	28.0	-0.14	28.9	28.8	0.02	X	X	X
White	77.1	72.9	0.10	47.6	45.0	0.06	76.1	75.0	0.03	67.1	70.0	-0.06
Female	75.2	71.4	0.08	77.4	77.8	-0.01	75.1	76.3	-0.03	70.1	71.0	-0.02
School Characteristics												
Eligible for free or reduced-price lunch	89.5	89.5	0.00	87.9	86.3	0.07	79.4	79.8	-0.02	84.9	85.0	-0.01
Non-White	94.3	95.0	0.03	95.1	94.5	0.06	69.3	68.6	0.04	89.0	87.8	0.08
ELL	25.6	25.6	0.00	16.0	15.5	0.02	15.4	15.0	0.03	5.3	4.7	0.09
Special education	15.2	15.3	-0.01	17.1	16.5	0.06	13.3	13.4	-0.03	10.8	10.9	-0.01
Reading proficiency	36.2	35.0	0.06	39.5	40.8	-0.07	35.2	34.3	0.06	57.0	57.7	-0.04
Mathematics proficiency	38.2	36.7	0.07	42.8	43.5	-0.04	34.1	33.3	0.06	53.9	56.8	-0.13
School size (average enrollment)	910	858	0.09	504	502	0.00	761	756	0.01	689	688	0.00

Note. X indicates that data on this variable were not available. *g* is the difference between groups in *SD* units. Differences were adjusted for each cohort. The sample size for some characteristics might be smaller because of missing data. Positive numbers for school characteristics represent the percentages of students in the school.

Source. AIR's analysis is based on data provided by the Chicago Public Schools, the District of Columbia Public Schools, Metropolitan Nashville Public Schools, and the Louisiana Department of Education.

Table E8. Baseline Characteristics of Teachers Included in the Analytic Samples for Teacher Instructional Practice in the Third Year of Teaching

Characteristics	Chicago (2012 Cohort)			District of Columbia (2010, 2011, and 2012 Cohorts Combined)			Nashville (2011 and 2012 Cohorts Combined)			New Orleans (2011 and 2012 Cohorts Combined)		
	Mean			Mean			Mean			Mean		
	Fellows (*n* = 40)	Comparison (*n* = 78)	*g*	Fellows (*n* = 137)	Comparison (*n* = 218)	*g*	Fellows (*n* = 44)	Comparison (*n* = 77)	*g*	Fellows (*n* = 16)	Comparison (*n* = 31)	*g*
Teacher Characteristics												
Age (years)	27.8	28.3	-0.12	28.4	29.1	-0.14	32.0	31.8	0.02	X	X	X
White	67.5	75.6	-0.18	56.1	46.4	0.19	65.9	71.4	-0.12	62.5	58.1	0.09
Female	75.0	71.8	0.07	78.3	80.7	-0.06	68.1	75.3	-0.16	81.4	77.4	0.09
School Characteristics												
Free or reduced-price lunch eligible	88.3	85.2	0.16	87.5	85.5	0.08	77.7	76.6	0.06	83.5	81.6	0.16
Non-White	92.9	92.0	0.07	93.1	92.3	0.05	66.1	65.1	0.06	86.6	84.9	0.09
ELL	20.6	19.8	0.04	16.9	15.0	0.09	13.8	12.6	0.09	8.4	9.5	-0.12
Special education	16.4	15.8	0.04	16.2	15.1	0.15	14.0	14.9	-0.10	9.4	9.3	0.04
Reading proficiency	35.5	37.6	-0.10	41.5	42.7	-0.06	37.8	38.6	-0.05	57.2	56.2	0.07
Mathematics proficiency	40.1	41.0	-0.04	46.4	46.5	0.00	39.9	40.8	-0.06	51.9	49.6	0.11
School size (average enrollment)	832	846	-0.03	519	524	-0.02	782	784	-0.01	781	859	-0.20

Note. X indicates that data on this variable were not available. *g* is the difference between groups in *SD* units. Differences were adjusted for each cohort. The sample size for some characteristics might be smaller because of missing data. Positive numbers for school characteristics represent the percentages of students in the school.

Source. AIR's analysis is based on data provided by the Chicago Public Schools, the District of Columbia Public Schools, Metropolitan Nashville Public Schools, and the Louisiana Department of Education.

AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW | Washington, DC 20007-3835 | 202.403.5000 | **www.air.org**

0846_5/17