SRI Education
A DIVISION OF SRI INTERNATIONAL

A Comprehensive Model of Teacher Induction: Implementation and Impact on Teachers and Students

Evaluation of the New Teacher Center's i3 Validation Grant, Final Report

Prepared by:

SRI Education

Viki M. Young
Rebecca Schmidt
Haiwen Wang
Lauren Cassidy
Katrina Laguarda

SRI International · 333 Ravenswood Avenue · Menlo Park, CA 94025-3493 · phone +1.650.859.2000 · www.sri.com

A Comprehensive Model of Teacher Induction: Implementation and Impact on Teachers and Students

Evaluation of the New Teacher Center's i3 Validation Grant, Final Report
(Appendix separate)

December 2017

Submitted to:

Ali Picucci

Vice President of Impact and Improvement
New Teacher Center

Prepared by:

SRI Education

Viki M. Young
Rebecca Schmidt
Haiwen Wang
Lauren Cassidy
Katrina Laguarda

Acknowledgments: The findings reported here culminated from the efforts of a large team over 5 years. The authors thank Marjorie Wechsler, Paul Hu, Hannah Cheever, Hannah Kistler, Andrew Ezekoye, Chi Nguyen, Bonnee Groover, Francine Biscocho, Juliet Tiffany-Morales, Jennifer Bland, Matt McCracken, Tiffany Hsieh, and Janelle Sands. The authors also thank Ali Picucci and the district teams at the New Teacher Center for their strong engagement with and feedback on the evaluation. Not least, the authors are grateful to the three partner sites—Broward County Public Schools, Chicago Public Schools, and Grant Wood Area Education Agency—for their participation in the evaluation.

Suggested citation: Young, V. M., Schmidt, R., Wang, H., Cassidy, L., & Laguarda, K. (2017, December). *A comprehensive model of teacher induction: Implementation and impact on teachers and students. Evaluation of the New Teacher Center's i3 Validation grant, final report*. Prepared for the New Teacher Center. Menlo Park, CA: SRI International.

CONTENTS

EXHIBITS

Teacher induction strategies aim to provide novice teachers with crucial supports as they first confront the realities of the classroom, shoring up essential management and instructional skills, improving retention in the profession, and ultimately bolstering student learning. The New Teacher Center (NTC) received an Investing in Innovation (i3) Validation grant in 2013 to implement its induction model in three sites: Broward County Public Schools (BCPS) in Florida, Chicago Public Schools (CPS) in Illinois, and the Grant Wood Area Education Agency (GWAEA), a consortium of rural districts in Iowa. Across the three sites, NTC trained full-time released mentors and served two cohorts of beginning teachers for 2 years each. Through the grant, NTC formalized four key components of its comprehensive induction model: (1) build the capacity of districts and school leaders to support the mentoring program, (2) select and assign full-time release mentors to caseloads of no more than 15 teachers each, (3) provide mentors more than 100 hours of intensive training through institutes and in-field support from lead coaches, and (4) provide regular, high-quality mentoring to first- and second-year teachers using a system of NTC-developed online formative assessment tools.

SRI Education conducted the evaluation of NTC's i3 Validation grant, examining the implementation and impact of NTC's induction model. The evaluation used a rigorous mixed-methods design to measure implementation fidelity and impact on teacher and student outcomes across the three participating sites. To account for different local contexts and needs, SRI used two methods to study impact: (1) randomized controlled trials (RCT) in BCPS and CPS with schools randomly assigned to NTC mentoring and control groups and (2) a quasi-experimental design (QED) in GWAEA. In each site, the evaluation team followed two cohorts of new teachers for 2 years each—Cohort 1 began teaching in 2013–14 and Cohort 2 in 2014–15—for a total implementation period of 3 years (2013–14 through 2015–16). The evaluation measured implementation across all 3 years and teacher and student impacts after teachers had participated in 2 years of induction.

Implementation Findings

Using teacher and mentor surveys, interviews, and NTC's online formative assessment system including a coaching log and tool data, the evaluation team annually examined each site's fidelity to the NTC key components. Additionally, SRI measured the extent to which the NTC model as implemented in treatment schools differed from the business-as-usual supports that new teachers received in control schools. The level of implementation fidelity and treatment-control differences helped indicate whether to expect an impact of the NTC induction model on teacher and student outcomes.

Implementation Fidelity

The fidelity of implementation analysis comprised four key program components: (1) NTC supports for the sites, (2) selection and assignment of high-quality mentors, (3) mentor development and accountability, and (4) provision of high-quality mentoring. Each component comprised three to eight indicators, each with defined thresholds for full (i.e., high), medium, and low implementation. Each site received a fidelity score for each indicator, and indicator-level scores were combined to create a site-level score for each key component. Each site's component scores were aggregated across all three sites for a program-level score.

The results across the 3 years showed high implementation fidelity for all sites. The sites improved their implementation of Component 3, mentor development and accountability, and Component 4, provision of high-quality mentoring, which had been scored as "medium" across the three sites in the first year (2013–14). This level of implementation fidelity in the first year was not

surprising, representing typical challenges of organizing the new induction strategy for newly selected and trained mentors and establishing relationships with schools and beginning teachers in that first year. In the second and third years (2014–15 and 2015–16), implementation fidelity was high on all key components, reflecting local focus and growth on the indicators central to NTC induction model.

Treatment–Control Contrast

On the annual surveys of teachers in treatment and control schools, treatment teachers consistently reported more robust induction supports than control teachers. Treatment teachers were more likely to report having a mentor than control teachers. Of those teachers who reported having mentors, treatment teachers met with mentors more frequently and for more time than control teachers and focused more on instruction during their meetings with mentors. Treatment teachers also rated the value of mentoring activities higher than control teachers and were more likely than control teachers to report that the induction supports they received helped them grow as teachers. These multiple measures of beginning teachers' induction experiences indicate that the NTC induction model indeed provided substantially different supports and experiences to treatment teachers from those reported by control teachers.

Teacher Impact Findings

The evaluation examined the extent to which the NTC induction model had an impact on teacher instructional practices and teacher retention in the RCT districts.[1]

Teacher Practice

To determine whether participating in the NTC induction model for 2 years resulted in better teaching practices, the evaluation team measured teacher practice outcomes through structured classroom observations using the Framework for Teaching (Danielson, 2013). Teachers of core subjects (mathematics, reading/English language arts, social studies, science, or self-contained elementary classrooms) in treatment and control schools were randomly selected and observed at two time points (baseline—at the start of their first year of teaching—and at the end of their second year of teaching).

The evaluation found no statistically significant differences between observed treatment and control teachers on the four measures of Domain 2: Classroom Management and the four measures of Domain 3: Instruction. However, because of attrition over time, the number of schools remaining in the analysis sample was relatively low, as was the number of teachers in each school, even when both cohorts and RCT districts were combined. The reduced sample size limited our ability to detect the effects of the NTC model on teacher practice using the Framework for Teaching, particularly if those effects were small or variability in practice among teachers was considerable.

Teacher Retention

Using district administrative data, SRI assessed the impacts of the NTC induction model on teachers' retention into their third year of teaching. Across both cohorts, 79 percent of treatment teachers and 78 percent of control teachers in the RCT district were retained; the difference was not statistically significant. The retention rates across both treatment and control teachers were lower than those found in a national sample of teachers beginning teaching in 2007–08, among whom 85 percent remained in teaching 3 years later (Gray & Taie, 2015). This difference raises the

[1] Teacher instructional practice could not be measured at baseline in the QED study because the comparison cohort began teaching before the start of the NTC grant. The teacher retention analysis from the QED site is a purely descriptive off-year comparison; therefore, we conducted only descriptive, not causal, analysis to inform NTC.

possibility that local factors and/or more recent trends may be influencing retention patterns that induction might not address.

Student Impact Findings

We examined whether the student achievement of teachers participating in the full NTC induction model for 2 years improved, specifically in English language arts (ELA) and mathematics among students in grades 4 through 8.[2] We used the Florida State Assessment (FSA) for BCPS and the Measures of Academic Progress (MAP) for CPS; CPS administered the MAP to have a consistent assessment that bridged the years during which Illinois switched state tests. For GWAEA, we used the state test, the Iowa Assessment.

RCT Sites [3]

The evaluation team found that NTC's induction program had overall positive effects on student achievement in ELA and mathematics in the two RCT districts (Exhibit ES-1).[4] The students in NTC-supported teachers' classroom for 1 year during the teachers' second year of support demonstrated higher achievement than students of teachers in the control group. In ELA, the average student achievement of teachers in the second year who participated in NTC induction for 2 years was approximately 0.05, compared with -0.04 for students of control teachers. This difference equals an effect size of 0.09 standard deviation ($p < .05$)—equivalent to moving from the 48th to the 52nd percentile—and represents the equivalent of approximately 2 to 3.5 additional months of learning, depending on the student's grade level (Lipsey, Puzio, Yun, Hebert, Steinka-Fry, Cole, et al., 2012).

In mathematics, students in grades 4 through 8 of teachers in the second year who participated in NTC induction for 2 years scored 0.15 standard deviation ($p < .01$) higher on average than students of control teachers. These impacts are equivalent to moving from the 46th to the 52nd percentile and represent the equivalent of approximately 2.4 to 4.5 additional months of learning, depending on the student's grade level.

Exhibit ES-1. Second-Year Impact on Student Achievement, Combined RCT Sites

	Adjusted Mean Test Scores		Difference	Sample Sizes		
Subject	Treatment	Control	(effect size)	Students	Teachers	Schools
ELA	0.05	-0.04	**0.09***	6,147	149	99
Math	0.06	-0.09	**0.15****	4,972	129	86

Note: The effect on student achievement is a 1-year effect as the districts provided current and prior achievement data annually but did not consistently provide identifiers to link students across the data sets given to researchers each year. The 1-year impact after 2 years of mentoring includes achievement in 2014–15 for Cohort 1 teachers and 2015–16 for Cohort 2 teachers.

Adjusted mean test scores are in standard deviation units.

* $p < .05$, ** $p < .01$

[2] Students in third grade take state assessments in Florida, Illinois, and Iowa. The third-grade scores serve as the measure of prior achievement for fourth-grade students. As the lowest tested grade, however, third-grade students do not have a measure of prior achievement and could not be included in the analysis. Fourth grade was the lowest grade that we could include in the sample.

[3] SRI released a findings brief in June 2017 with the student achievement results from the RCTs. Schmidt, R., Young., Cassidy, L., Wang, H., & Laguarda, K. (2017, June). *Impact of the New Teacher Center's new teacher induction model on teachers and students.* Menlo Park, CA: SRI International. https://www.sri.com/work/publications/impact-new-teacher-centers-new-teacher-induction-model-teachers-and-students

[4] District results varied; see Appendix F for methods and district results.

QED Site

In the quasi-experimental study, SRI used a difference-in-differences approach to estimate the impact of participating in the 2-year NTC induction program. The study compared the difference in student achievement between beginning teachers who started teaching in 2013–14 and received NTC induction support for 2 years and a cohort of comparison beginning teachers who started teaching in 2012–13 and did not receive NTC induction support with the difference in the student achievement of veteran teachers in the same years.[5]

The impact estimate for teachers beginning teaching in 2013–14 and in their second year of induction support was not statistically significant, suggesting no detected NTC impact on this cohort of teachers in the QED site. However, the sample size of beginning teachers teaching ELA or mathematics in grades 4 through 8 that resulted from the participating districts' hiring patterns and testing schedules was very small, with 8 comparison and 19 treatment teachers in the ELA analysis and 7 and 23, respectively, in the mathematics analysis. The QED was extremely constrained in being able to detect any effects. As a result, the QED results are inconclusive—we do not know whether the NTC induction model had an impact in the QED site and these results should be interpreted with caution.

Conclusions and Implications

The high implementation fidelity levels and contrasts in induction experiences between treatment and control teachers indicate that the NTC induction model can be implemented well in a range of district contexts and even during times of budget cutbacks, as was the case in CPS. NTC induction did not yield differences in teacher practice as measured through classroom observations, although the sample sizes were small, and in teacher retention rates between treatment and control groups.

The positive student impacts in the RCT sites suggest that the NTC induction model can improve the ELA and mathematics achievement of students in beginning teachers' classrooms. The QED using a differences-in-differences approach did not bear out positive impacts on student outcomes, but it was limited by the small sample size and we do not know statistically whether the NTC induction model had an impact in the QED site.

The mixed results of positive impact on student outcomes but not on teacher practices warrants further investigation. The lack of impact on teacher practices was most likely due to attrition and small sample size. In addition, it is possible that the measures of teacher practice were not fine-grained enough to capture the nature of NTC effects on instruction.

Building on these results under an i3 Scale Up grant, NTC is currently implementing its model in five urban districts across the country and SRI is conducting RCTs in each district. Although NTC successfully achieved high implementation fidelity under the i3 Validation grant, scaling up to more districts and more diverse contexts necessitated adaptations to enhance sustainability and applicability. The evaluation of the i3 Scale Up grant will examine further whether and to what extent the NTC induction model incorporating certain adaptations, such as school-based and part-time mentors and classroom video tools, can achieve high implementation fidelity in larger and more diverse district settings. It will also explore whether, across these varying contexts, the NTC induction model has positive effects on teacher practice, teacher retention, and student achievement.

---

[5] When comparing different cohorts of new teachers/their students across years, the intervention and comparison conditions are completely aligned with different time periods, and the estimated impact is confounded with policy or environmental changes from one year to the next that might have affected achievement. By including the veteran teachers from each time period as extra comparison groups for the intervention and comparison new teachers, respectively, this difference-in-differences design attempts to address this issue by controlling for changes that may have occurred between time periods, therefore eliminating the confounding with time.

INTRODUCTION

Few professions demand that from their first days on the job, novices perform the same duties at the same level as seasoned veterans. The teaching profession routinely does. In the first years of their careers, teachers experiment—and often struggle alone—with managing student behavior, mastering the curriculum, engaging students in their own learning, pacing activities, and attending to differences in how students learn and to their diverse academic and social needs. Reformers have argued that during their formative years, new teachers also set the foundation for habits and dispositions that persist through their careers (Snyder & Bristol, 2015). Concerns about instructional quality and students' resulting academic performance (Hanushek, 1992; National Commission on Teaching and America's Future, 1997, 2016; Sanders & Rivers, 1996), as well as turnover (Smith & Ingersoll, 2004), further underscore the need for robust induction supports for beginning teachers.

The New Teacher Center (NTC), headquartered in Santa Cruz, California, was at the forefront of developing comprehensive induction strategies as California put in place a statewide induction policy, Beginning Teacher Support and Assessment (BTSA), in the late 1990s. Over the years, NTC has refined a comprehensive mentor-based induction model and served beginning teachers in districts across the country. NTC received an Investing in Innovation (i3) Validation grant in 2013 to implement its induction model in three sites: Broward County Public Schools in Florida, Chicago Public Schools in Illinois, and the Grant Wood Area Education Agency, a consortium of rural districts in Iowa. Through the grant, NTC trained full-time released mentors and served two cohorts of beginning teachers for 2 years each across the three sites.

SRI Education conducted the evaluation of NTC's induction model under the i3 Validation grant. The evaluation featured a rigorous mixed-methods design to measure implementation fidelity and impact on teacher and student outcomes in the three participating sites. To accommodate local program needs, SRI conducted randomized controlled trials in Broward County Public Schools and Chicago Public Schools and a quasi-experimental study in Grant Wood Area Education Agency. The evaluation team used multiple measures to capture implementation fidelity and provide timely feedback to NTC and to the sites and to inform the outcomes analyses.

This report begins with an overview of NTC's induction program, the study design, and evaluation activities. It then presents results from the implementation study, gleaned from analyses of implementation fidelity and the degree of contrast in mentoring that teachers in the treatment group received compared with that of teachers in the control group. Finally, the report examines the effect of the NTC induction model on teacher outcomes, including retention and classroom practice, and on student achievement in English language arts and mathematics after 2 years of induction support for teachers.[6] A separate published file includes this final report and comprehensive methods appendices with supporting tables.[7]

Program Description

NTC has long worked with district partners to implement a high-quality mentoring and induction program. Under the i3 Validation grant, NTC formalized key components of its induction model. NTC provides professional development, research-based resources, and online formative

[6] SRI released a findings brief in June 2017 with the student achievement results from the RCTs. Schmidt, R., Young., Cassidy, L., Wang, H., & Laguarda, K. (2017, June). *Impact of the New Teacher Center's new teacher induction model on teachers and students.* Menlo Park, CA: SRI International. https://www.sri.com/work/publications/impact-new-teacher-centers-new-teacher-induction-model-teachers-and-students

[7] https://www.sri.com/work/publications/comprehensive-model-teacher-induction-implementation-and-impact-teachers

assessment tools for beginning teachers, mentors, and school leaders, as well as technical assistance and capacity building for program leaders.

Logic Model

As depicted in the logic model (Exhibit 1), the NTC induction model featured carefully selected full-time mentors housed in district-level teacher development offices. These mentors received more than 100 hours of training annually from NTC program staff, during institutes, and through in-field support from local induction program leaders and lead coaches. The mentors, who were supervised centrally, supported first- and second-year teachers in multiple schools at a ratio of 15 beginning teachers to 1 mentor. New teachers received 2 years of coaching, meeting with their assigned mentors weekly for a minimum of 180 minutes per month. Mentors and teachers worked through a system of NTC-developed online formative assessments, including tools to guide observation cycles and to develop teachers' skills in planning lessons and analyzing student work. These features are consistent with Ingersoll and Strong's (2011) meta-analysis identifying characteristics of high-quality induction, including mentoring, common and regular planning time with the mentor, and mentor training, similarly cited by Hobson et al. (2009).[8] The three components of NTC induction—mentor selection and assignment, mentor development and accountability, and high-quality mentoring—were intended to improve instructional effectiveness, increase teacher retention, and ultimately improve student achievement (Exhibit 1, far right boxes). Supporting these three components are NTC's efforts to build district capacity to sustain teacher induction over the long term and resources, tools, and convenings led by NTC's national office (Exhibit 1, far left box). Key district and school conditions provided the context for mentors' opportunities to work with beginning teachers and shaped teachers' evolving practice (Exhibit 1, top box).

Several elements distinguished the NTC induction model from traditional district mentoring programs. The teacher to mentor ratio was intentionally low to enable mentors to work with new teachers frequently, intensively during each meeting, and consistently during the school year. The induction model encompassed the first 2 years in the classroom, when novices need to rapidly master classroom management and pedagogical skills and build the foundation for a sustainable teaching career. It is at this time that they are also at high risk of leaving the profession, which sustained induction support is intended to mitigate (Smith & Ingersoll, 2004).

Comprehensive mentor training and a system of formative assessment tools shaped the content and quality of mentor-teacher interactions under the NTC induction model. Mentors received a series of 12 professional learning days over 2 years.[9] These Mentor Academies introduced the mentors to key tenets of coaching (e.g., taking a collaborative stance with beginning teachers, focusing on equitable instruction), mentoring skills (e.g., observing and giving feedback), and formative assessment tools intended to aid mentors in enacting those tenets in their work with new teachers. Mentor Academies developed mentors' expertise in identifying effective teacher practice, using data to inform instruction, creating classroom conditions to foster equitable learning, supporting language development, and differentiating instruction for diverse learners.[10] The training focused mentors' interactions with beginning teachers on instruction, in contrast to mentoring that often provides logistical and emotional support for teachers and lacks sufficient emphasis on the instruction demanded by high content standards (Wang & Odell, 2002). Monthly Mentor Forums provided additional opportunity for mentors to reflect on how effectively they used

8 NTC induction was one of the programs in the 15 studies Ingersoll and Strong (2011) reviewed.

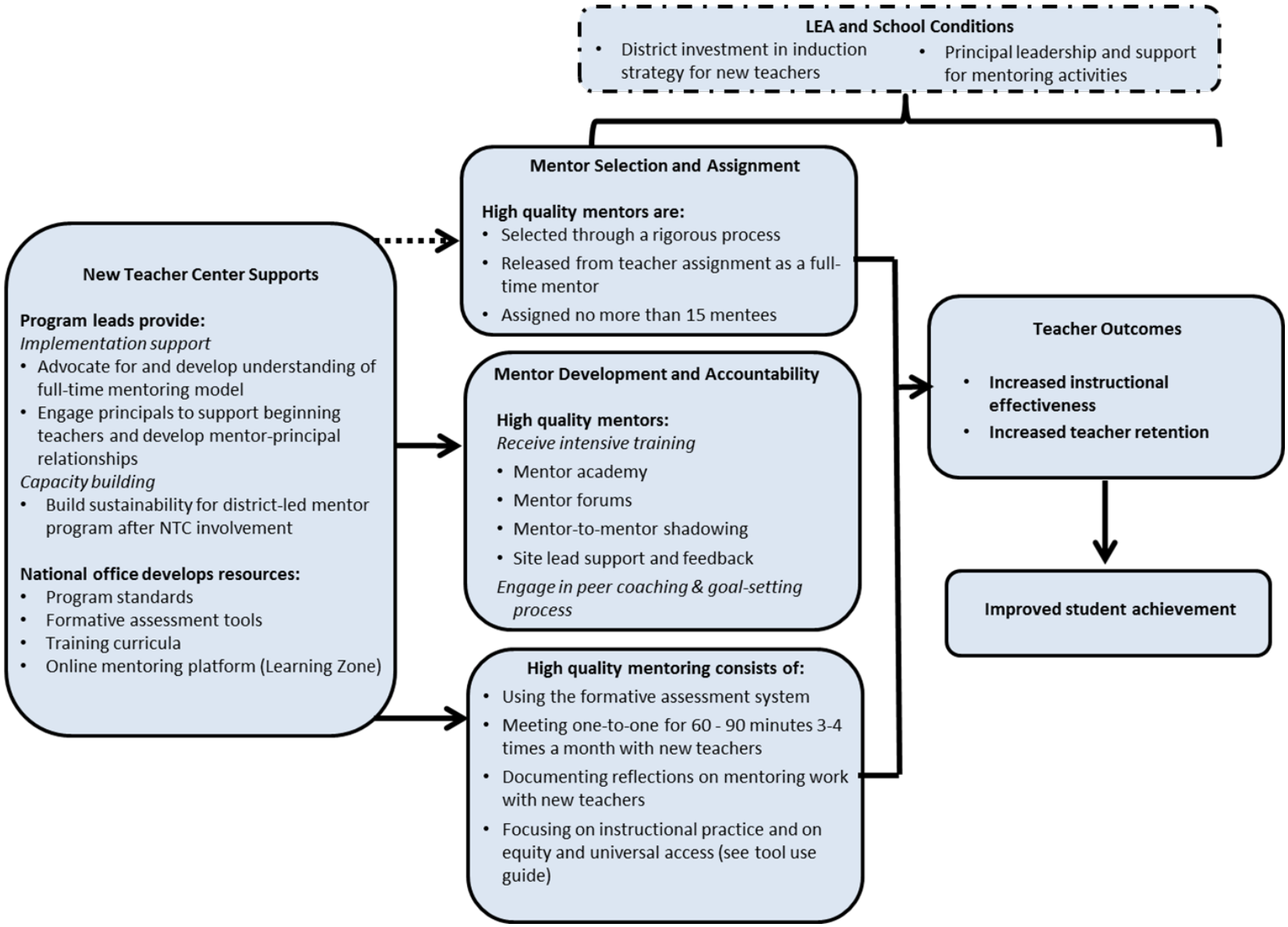
9 NTC has since revised its mentor training and now offers the 8-day Professional Learning Series (PLS) for Mentors, as well as Mentor Forums and in-field coaching.

10 See, for example, https://newteachercenter.org/wp-content/uploads/Professional-Learning-Series-Product-Sheet.pdf

tools aligned with specific instructional strategies, to raise questions about how to use particular tools, and to brainstorm and problem-solve with fellow mentors on how to meet specific teachers' needs or common needs across many teachers. In-field coaching, where a lead coach observed a mentor working with a beginning teacher and provided the mentor with feedback on that interaction, further supported mentors in refining their coaching practice and instructional support.

The formative assessment system was a comprehensive set of NTC-developed, instructionally focused tools and protocols. The tools support mentors in structuring their mentoring sessions with beginning teachers and ensure that the conversation and activities drove toward a specific instructionally focused objective for that coaching session. NTC highlighted lesson planning, guiding purposeful classroom observations and conferencing, and analyzing student work as high-leverage skills with aligned tools. All tools resided in NTC's online system, Learning Zone, in which mentors or new teachers entered the content as they worked through a tool. Learning Zone provided summary data on mentoring frequency and duration, numbers of tools mentors used, and which tools they used during each mentor-new teacher meeting. Additionally, mentors and teachers together could reference their prior work with the tools to understand their progress. NTC provided Learning Zone data monthly to the sites as ongoing feedback on the extent to which mentors were meeting expectations in mentoring frequency, duration, and tool use.

Exhibit 1. Logic Model for the New Teacher Center i3 Validation Grant



Implementing Sites

Three sites participated in the study—Broward County Public Schools (BCPS) in Florida, Chicago Public Schools (CPS) in Illinois, and Grant Wood Area Education Agency (GWAEA), a consortium of mainly rural districts in Iowa. Through the grant, NTC provided funds to hire full-time released mentors in each site and site-based lead coaches responsible for mentor training and support at their respective sites. The grant also covered district administrators' time in overseeing program implementation and data collection at the sites. In each site, the evaluation followed two cohorts of new teachers for 2 years each—Cohort 1 began teaching in 2013–14 and Cohort 2 in 2014–15—for a total implementation period of 3 years (2013–14 through 2015–16). Designated NTC staff members served as site-specific client liaisons to communicate with site executives and administrators, integrate induction into the sites' overall teacher development strategy, and build local capacity to sustain induction over the long term.

In the urban sites (BCPS and CPS), the grant funds provided the capacity to serve new teachers in a subset of schools, capped at 15 new teachers for 2 years each per funded mentor. Fifteen mentors were funded throughout the implementation period in CPS; BCPS had 9 mentors in the first year of implementation, and the number increased to 15 in the second year. In GWAEA, the participating districts were relatively small and had fewer schools, so they pooled resources to support mentors to work with beginning teachers across the districts. Any given district in the consortium had few new teachers each year, so the funded mentors in GWAEA served all new teachers during the implementation period.

These differences in program implementation necessitated two approaches to evaluating effectiveness. Because the number of new teachers in BCPS and CPS exceeded the capacity for induction support under the grant, we were able to use randomized controlled trials (RCTs), with random assignment as the mechanism for allocating NTC induction support to schools. In GWAEA, random assignment was not practical, so we adopted a quasi-experimental design (QED) to determine the effectiveness of NTC mentoring on teacher and student outcomes.

Within each site, the characteristics of the schools in the study generally reflected those of the district or the consortium, in the case of GWAEA, overall, with some slight differences. On average, the CPS study schools had higher proportions of English learners than the district overall and had higher school ratings. The BCPS study schools had slightly higher average percentages of students eligible for free or reduced-price lunch and racial/ethnic minority students than the district overall (Exhibit 2). In GWAEA, study schools tended to have slightly higher proportions of students eligible for free or reduced-priced lunch, and treatment schools in the student outcomes analysis were lower performing on average in English language arts (ELA) and mathematics compared with the state (Exhibit 3).

Exhibit 2. Average District and Study Sample Characteristics at the Time of Random Assignment, RCT Districts

		No. of Schools[a]	No. of Teachers in Study	School Rating	% English Learners	% Special Education	% Free or Reduced-Price Lunch Eligible	% Minority
BCPS	Treatment	43	193	2.2	10	13	71	81
	Control	44	148	2.6	11	16	68	79
	District	213	--	2.7	12	13	66	77
CPS	Treatment	65	149	2.4	22	13	84	92
	Control	75	139	2.5	24	12	84	88
	District	536	--	2.1	14	14	84	91

Note: "School Rating" refers to the state report card of quality ratings applied to each school. At the time of random assignment, BCPS assigned all schools a letter grade (A to F), and CPS used whole numbers between 1 and 3. These ratings were put on a common scale where 0 = F in BCPS and 3 in CPS, 1 = D in BCPS, 2 = C in BCPS and 2 in CPS, 3 = B in BCPS, and 4 = A in BCPS and 1 in CPS.

[a] School count does not include high schools.

Source: http://cps.edu/SchoolData/Pages/SchoolData.aspx (CPS); http://www.broward.k12.fl.us/dsa/counts/1213/20DayCount1213.shtml; and schoolgrades.fldoe.org/xls/1213/SGbasic_2013.xls (BCPS)

Exhibit 3. Average Consortium and Study Sample Characteristics, Year Before Intervention, QED Site

	No. of Schools	No. of Teachers in Study[a]	No. of New Teachers in Study	% Passing ELA	% Passing Math	% English Learners	% Special Education	% Free or Reduced-Price Lunch Eligible	% Minority
Treatment sample in student outcomes analysis	24	159	34	67	69	2	16	39	13
Control sample in student outcomes analysis	14	95	16	73	77	1	14	35	10
Treatment group with new teachers	57	112	112	77	72	2	14	37	12
Control group with new teachers	51	113	113	78	73	1	12	36	10
Consortium[b]	88	NA	NA	NA	NA	1	--	33	11

Note: Data for Cohort 1 teachers and their schools only.

State-level passing rate was 70.6% for reading and 76.8% for math across grades 3 through 8.

[a] Number of teachers in study includes veteran comparison teachers.

[b] Consortium data do not include high schools. Special education data were not available at the consortium level.

Source: https://portal.ed.iowa.gov/iowalandingpage/Landing.aspx, and additional individual school-level data from Iowa Department of Education.

Study Design Summary

The i3 Validation grant requires that evaluations include an implementation study to examine implementation fidelity in the participating sites and an impact study to determine whether the program as implemented had an impact on student outcomes and relevant teacher outcomes.

Implementation Study

In the implementation study, we examined the extent to which each site implemented the full induction program as described in the NTC logic model (Exhibit 1) and aggregated implementation fidelity scores across the three sites, addressing the research question: *What is the level of implementation fidelity to the NTC model in the three participating sites?*

Implementation Fidelity

We measured fidelity of the implementation of the four key program components depicted in the logic model: (1) NTC supports for the sites, (2) selection and assignment of high-quality mentors, (3) mentor development and accountability, and (4) provision of high-quality mentoring. Each component comprised three to eight indicators, which we developed in collaboration with NTC staff. Some indicators were measured at the site level, others at the individual level (e.g., by principal, mentor, or new teacher). For each indicator, SRI worked with NTC staff to set thresholds for full (i.e., high), medium, and low implementation. Each site received a fidelity score for each indicator, and the scores were combined to create a site-level score for each key component using the following rules:

- High implementation fidelity—60 percent or more of the indicators were scored as high, and no more than 20 percent of the indicators were scored as low.
- Medium implementation fidelity—Individual indicator scores did not reach the threshold for high fidelity, and less than 50 percent of indicators were scored as low.
- Low implementation fidelity—50 percent or more of the indicators were scored as low.

At the program level (i.e., across all sites in the study), NTC achieved implementation with fidelity under each key component if at least two sites achieved high implementation and no site achieved low implementation.

Multiple data sources were required to measure the constituent indicators for each component (Exhibit 4); we collected implementation data and calculated fidelity measures annually.

Exhibit 4. Data Sources for Implementation Analysis of Key Components

Component	No. of Indicators	Data Sources
NTC supports	Year 1: 7 Years 2 & 3: 8	Attendance log at half-day principal training; logs of one-on-one meetings between site leads and principals; copies of program standards, formative assessment tools and mentor training materials; Learning Zone data
Mentor selection and assignment	3	Mentor application materials; mentor survey; rosters of teacher assignments to mentors
Mentor development and accountability	7	Attendance log at mentor academies and mentor forums; logs of mentor-to-mentor shadowing, meetings between site leads and mentors, peer coaching, and goal setting
Provision of high-quality mentoring	5	Learning Zone data; teacher survey

Appendix A contains a full matrix that defines the specific implementation measures, data sources, and fidelity scores across the three sites for the 3 implementation years.

Teacher Survey

Annual spring surveys of the new teachers being served by NTC-trained mentors and their counterparts in control schools provided measures of the extent to which induction activities fundamental to the NTC induction model differed between treatment and control schools (treatment-control contrast). The survey contained items about the mentoring new teachers received (frequency and duration of meetings, focus of mentors' work with new teachers), other kinds of induction supports, school environment, and beginning teachers' self-evaluation. To understand treatment-control contrast across the 3 years of the program, we analyzed the surveys administered between spring 2013 and spring 2016. The analysis combined teachers across cohorts in their first and second years of teaching as shown in Exhibit 5, which matched the impact analyses combining both teacher cohorts in the RCT sites.

Exhibit 5. School Year by Cohort and Years of Experience

	Years of Teaching Experience	
New Teacher Cohort	**1**	**2**
1	2013–14	2014–15
2	2014–15	2015–16

A total of 860 first-year teachers across both cohorts and 660 second-year teachers across both cohorts in all three sites responded to the survey. Overall, 90 percent of treatment teachers and 64 percent of control teachers responded to the survey (Exhibit 6). The survey sample sizes and response rates by site are presented in Appendix B.

Exhibit 6. Overall Survey Response Rates by Years of Teaching Experience, Combined Cohorts

	Years of Experience—Both Cohorts								
	Year 1[a]			Year 2			Overall		
Site	Treatment	Control[b]	Subtotal	Treatment	Control	Subtotal	Treatment	Control	Total
Surveyed	608	417	1025	518	373	891	1126	790	1916
Responded	563	297	860	450	210	660	1013	507	1520
Response rate	93%	71%	84%	87%	56%	74%	90%	64%	79%

Source: NTC New Teacher Survey, spring 2013–2016.

[a] Attrition data were not available for the Year 1 calculations.

[b] The control group was administered the teacher survey 1 year before the treatment group in GWAEA.

Impact Studies

As noted, because of differences in local contexts we used two approaches to examine the impact of NTC's induction model, RCTs in BCPS and CPS and a QED in GWAEA.

Randomized Controlled Trials in BCPS and CPS

The RCTs in CPS and BCPS featured school-level random assignment to estimate the impact of the NTC model on teacher and student outcomes. In both districts, we randomly assigned a sample of participating schools employing beginning teachers in summer 2013, before NTC began serving the new teachers. The schools in each district were blocked on grades served and the most relevant

local factor—geographic area in CPS and Teacher Incentive Fund status in BCPS. All schools were assigned and all teachers identified by October 1. Within each block, schools were assigned to the NTC program or to the usual district supports for new teachers until the target number of beginning teachers was reached (before October 1). In the second year (2014–15), all incoming first-year teachers in treatment schools were added to the treatment group and all incoming first-year teachers in control schools were added to the control group. To reach the target number of teachers served, previously unassigned schools in each district were assigned in summer 2014 following the established blocks if they employed beginning teachers before October 1. See Appendix C for assignment details.

In both sites, NTC served all new teachers in treatment schools unless they were served by other programs with induction support, such as Teach For America. Teachers covered under other induction programs were excluded from both the treatment and control conditions.

The RCTs in BCPS and CPS address the following research questions about NTC impacts:

1. Does participating in the full NTC induction model for 2 years result in better practices on eight components of teaching? *(Confirmatory)*
2. Does participating in the full NTC induction model for 2 years result in improved student achievement in reading and math among students in grades 4–8? *(Confirmatory)*
3. Does participating in the full NTC induction model result in improved teacher retention after 2 years? *(Exploratory)*[11]

Quasi-experimental Design in GWAEA

Sixteen districts in the GWAEA consortium agreed to participate in the study in 2013–14. Because all beginning teachers in the GWAEA participating districts began receiving NTC mentoring in 2013–14, we could not use random assignment to study impact. Instead, we used a quasi-experimental difference-in-differences approach to estimate impact on student achievement. That is, we compared beginning teachers in GWAEA participating districts who began teaching in 2012–13 and did not receive NTC mentoring with beginning teachers in the same districts who began teaching in 2013–14, when NTC mentoring was offered to all new teachers, adjusting for differences between veteran teachers in those years. We included veteran teachers in the analysis because when comparing different cohorts of new teachers/their students across years, the intervention and comparison conditions are completely aligned with different time periods, and the estimated impact is confounded with policy or environmental changes from one year to the next that might have affected achievement. By including the veteran teachers from each time period as extra comparison groups for the intervention and comparison new teachers, respectively, this difference-in-differences design attempted to address this issue by controlling for changes that may have occurred between time periods, therefore eliminating the confounding with time.[12] Therefore, to meet this standard, we could compare the differences only between the first cohort of NTC-

[11] Confirmatory questions are those related to impact that the evaluation team defines a priori, before examining outcome data. Exploratory questions are not specified in advance, may be more responsive to program developers' information needs, and may change or be developed to investigate questions that arise after seeing findings.

[12] From communication between the National Evaluation of i3 staff and the SRI research team, this difference-in-differences approach has better validity and is more likely to meet the What Works Clearinghouse (WWC) standards when treatment and comparison groups are no more than 1 year apart, although WWC changed the standards for difference-in-differences approaches in 2014, after we had designed the impact study for GWAEA and the site had begun serving new teachers. SRI conducted student outcomes analysis for GWAEA teachers in the second cohort to inform NTC about its program. The results are in Appendix H and are not intended for WWC review.

served teachers in the 16 districts that volunteered to participate in the study in 2013–14 and the beginning teachers in 2012–13 in these districts who were not served by NTC.

The impact study in GWAEA addresses the research question: *Does participating in the full NTC induction model for 2 years result in improved student achievement in reading and math among students in grades 4–8?*

Data Collection Activities

Over the course of the study, we collected and analyzed data from multiple and varied sources to gain a comprehensive view of implementation and impact. Exhibit 7 indicates the data sources by purpose. Exhibits 8 and 9 detail the data collection activities by cohort, implementation year, and type of school for the two impact studies. For the implementation study, we analyzed Learning Zone data on mentoring activities for each year of implementation (2013–14 to 2015–16), NTC-administered teacher and mentor survey data, school- and district-level interviews, and extant data. For the impact studies, we conducted and analyzed classroom observations and teacher retention data from the RCT sites and collected and analyzed student-level achievement and demographic data from all participating sites.

Exhibit 7. Data Sources and Their Purposes

	Purpose					
Data Source	Randomly Assign Schools	Teacher Eligibility	Teacher Outcomes Analysis	Student Outcomes Analysis	Sensitivity and Follow-up Analyses	Implementation and Treatment-Control Contrast
School demographic and achievement data	✓		✓	✓	✓	✓
Human resources data		✓	✓	✓	✓	
Teacher observations			✓		✓	
Teacher, mentor, school leader, and district leader interviews						✓
Teacher and mentor surveys					✓	✓
Student demographics and achievement data				✓	✓	

Exhibit 8. Data Collection Activities by Cohort and Year for BCPS and CPS

Random Assignment Group	2013-14 Cohort 1		2014-15 Cohort 1		2014-15 Cohort 2		2015-16 Cohort 1		2015-16 Cohort 2		2016-17 Cohort 2	
	T	C	T	C	T	C	T	C	T	C	T	C
Learning Zone	✓		✓		✓				✓			
Teacher survey	✓	✓	✓	✓	✓	✓			✓	✓		
Mentor survey	✓		✓		✓				✓			
Interviews			✓		✓				✓	✓		
Classroom observations	✓	✓	✓	✓	✓	✓			✓	✓		
Student achievement & demographic data	✓	✓	✓	✓	✓	✓			✓	✓		
Human resources data	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Cohort 1 = Teachers beginning teaching in 2013–14.
Cohort 2 = Teachers beginning teaching in 2014–15.
T = treatment group.
C = control group.

Exhibit 9. Data Collection by Cohort and Year for GWAEA

	2012–13			2013–14			2014–15			2015–16		
	Com	C1	C2	Com	C1	C2	Com	C1	C2	Com	C1	C2
Implementation Data												
Learning Zone		--	--		✓	--		✓	✓	--	--	✓
Teacher survey	✓	--	--	✓	✓	--		✓	✓	--	--	✓
Mentor survey		--	--		✓	--		✓	✓	--	--	✓
Interviews		--	--			--		✓		--	--	✓
Impact Data												
Classroom observations	✓	--	--	✓		--		✓		--	--	
Student achievement & demographic data[a]	✓	--	--	✓		--	✓	✓		--	--	✓
Human resources data	✓	--	--	✓	✓	--	✓	✓	✓		✓	✓

Comparison (Com) = Cohort of teachers beginning teaching in 2012–13.
C1 = Cohort 1 treatment teachers beginning teaching in 2013–14.
C2 = Cohort 2 treatment teachers beginning teaching in 2014–15.
[a] We also collected student achievement and demographic data for veteran teachers in the same years as the comparison cohort and Cohorts 1 and 2 for the difference-in-differences analysis.

Broward County Public Schools in Florida, Chicago Public Schools, and the Grant Wood Area Education Agency in Iowa—the participating sites—had differing contexts that influenced how the NTC induction model was implemented locally. Therefore, understanding the extent to which each site was able to implement the key components of the NTC model given varying local conditions was the first step in establishing whether the overall model was implemented to a level of fidelity that would predict an impact on the target teacher and student outcomes.

Throughout the study, we provided NTC with information on implementation as we completed data analysis to help program leaders and sites identify specific areas for improvement, such as barriers to mentors being able to meet with their assigned beginning teachers regularly or supports that mentors and teachers might need to use the formative assessment tools well. In addition to measuring implementation fidelity and supporting program improvement, we collected and analyzed data from beginning teachers in treatment and control schools to understand any differences in their induction experiences. These treatment-control differences signaled whether to expect an impact of the NTC induction model on teacher and student outcomes compared with the status quo supports that control teachers received in their first and second years of teaching. This chapter presents the implementation fidelity data for all 3 years of implementation and discusses the evidence on the extent to which treatment and control teachers' induction experiences differed.

Implementation Fidelity

The NTC induction logic model identified four key components for which we measured implementation fidelity:

1. NTC supports—Eight indicators of the supports NTC provided in launching the program in each site. One indicator (capacity-building by site leads) was measured only in years 2 and 3.
2. Mentor selection and assignment—Three indicators addressing mentor hiring and allocation to new teachers.
3. Mentor development and accountability—Seven indicators of the site-level supports and training for mentors.
4. Provision of high-quality mentoring—Five indicators reflecting the joint activities mentors and beginning teachers engaged in and teachers' perceptions of the quality of their mentoring experience.

Exhibit 10 provides the scores for each component, aggregated across the three sites for a program-level score. The results pertain to the overall program serving both first- and second-year teachers in 2013–14, 2014–15, and 2015–16. Appendix A presents the full matrix that defines the specific measures, data sources, and fidelity scores across the sample for the 3 study years.

Exhibit 10. Implementation Fidelity by Key Component

Key Component	Number of Indicators	Year 1 (2013–14) *Cohort 1, first year teaching*		Year 2 (2014–15) *Cohort 1, second year teaching* *Cohort 2, first year teaching*		Year 3 (2015–16) *Cohort 2, second year teaching*	
		Number of Sites Meeting Fidelity Threshold	Fidelity, Program Level (High/ Medium/ Low)	Number of Sites Meeting Fidelity Threshold	Fidelity, Program Level (High/ Medium/ Low)	Number of Sites Meeting Fidelity Threshold	Fidelity, Program Level (High/ Medium/ Low)
1. New Teacher Center supports	8	High: **3** Medium: **0** Low: **0**	**High**	High: **3** Medium: **0** Low: **0**	**High**	High: **2** Medium: **1** Low: **0**	**High**
2. Mentor selection and assignment	3	High: **3** Medium: **0** Low: **0**	**High**	High: **3** Medium: **0** Low: **0**	**High**	High: **3** Medium: **0** Low: **0**	**High**
3. Mentor development and accountability	7	High: **0** Medium: **3** Low: **0**	**Medium**	High: **2** Medium: **1** Low: **0**	**High**	High: **2** Medium: **1** Low: **0**	**High**
4. Provision of high-quality mentoring	5	High: **1** Medium: **2** Low: **0**	**Medium**	High: **3** Medium: **0** Low: **0**	**High**	High: **3** Medium: **0** Low: **0**	**High**

Note: "High" implementation at program level (three sites) for each key component = at least two sites scored high and none scored low for that component.

The results across 3 years show that implementation fidelity was generally high. However, the implementation fidelity of Component 3, mentor development and accountability, and Component 4, provision of high-quality mentoring, was lower in the first year (2013–14), scoring "medium" across the three sites. This level of implementation fidelity was not surprising for the first year of program launch, when all sites were starting a brand-new program.

In the second and third years (2014–15 and 2015–16), implementation fidelity was high on all key components, showing local focus and growth on the indicators that define the NTC induction model. For Component 3 in particular, mentors in all three sites were more consistent in attending Mentor Academies and Mentor Forums. All sites met the definition for high implementation fidelity for that indicator (80 percent of mentors attended forums and academies offered). All three sites also improved in having mentors meet regularly with beginning teachers and using the NTC formative assessment tools during their meetings with mentees, driving improvement under Component 4.

Despite generally high implementation, some inconsistencies in the first three components in the third year (2015–16) reflected local challenges that were outside NTC's control. For example, intensifying local budget constraints affected mentor workloads and retention and created uncertainty about continued funding for the induction program beyond the grant period. In the last year of implementation, the sites were less consistent in the following indicators under each component:

- Component 1, New Teacher Center supports—Engaging principals was less consistent, with two districts not meeting the high definition for the indicator relating to annual one-on-one meetings between site leads and principals.
- Component 2, mentor selection and assignment—Maintaining a caseload of 15 teachers to each coach was less consistent. Two districts met this caseload ratio for more than 90 percent of mentors in the second year, but in the third year two districts were able to meet this caseload ratio for only over 80 percent of mentors.
- Component 3, mentor development and accountability—Mentor-to-mentor shadowing and mentors' receiving feedback from site leads was less consistent, reflecting the general pressure on mentors' and program staff's time.

Despite these challenges, teachers' mentoring experiences remained of high quality. On Component 4, provision of high-quality mentoring, all sites met the standards for frequency and intensity of mentoring and for focus on instruction during mentoring. Program staff members and mentors placed a primacy on serving beginning teachers and preserving the level of attention they needed, choosing to scale back on some of the mentor supports and principal engagement instead.

Contrast in Induction Supports Between Treatment and Control Schools

Beyond establishing the level of implementation fidelity in treatment schools, putting the impact of the NTC model in perspective requires an understanding of any differences in the comprehensive induction that NTC provides in treatment schools with business-as-usual supports that districts and schools normally provide beginning teachers in control schools. Over the course of a multiyear study and particularly for a 2-year intervention such as the NTC induction model, the scope and the quality of status quo induction supports may change as implementing sites seek to improve their own programs, so the differences in induction experiences between teachers in treatment and control schools may fluctuate over time.

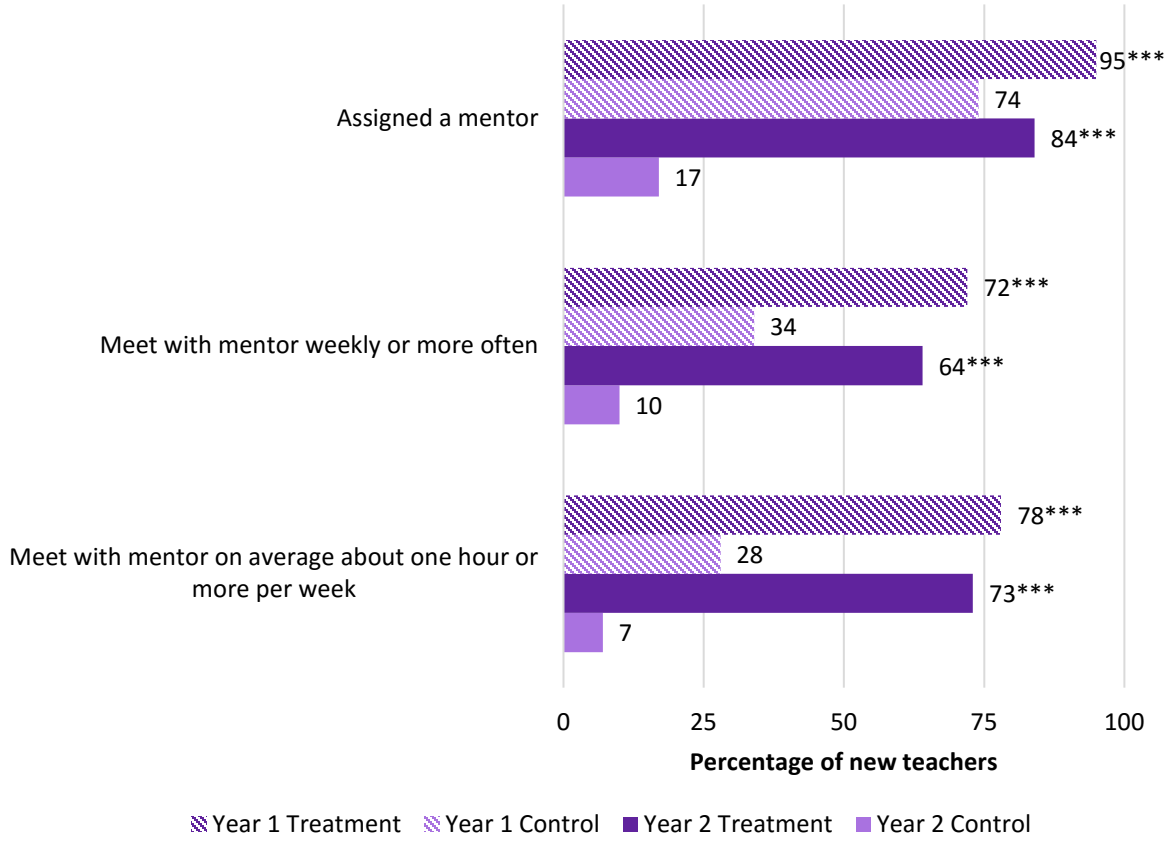
Characterizing this treatment-control contrast provides important context for understanding any impact of NTC induction on teachers' practice and students' achievement. Based on NTC's logic

model, impacts would be expected to be greater in sites where NTC services are very different in scope and quality from the districts' status quo approach to induction. Similarly, where we find variations in the treatment-control contrast, we might also expect to see variations in impact.

To parallel and more accurately inform the impact findings, we conducted the analyses of sites in the RCT and QED separately. All results discussed here were statistically significant at the $p < .05$ level. Survey scale items are listed in Appendix B.

The survey results showed consistent differences between treatment and control teachers. Overall, treatment teachers were more likely to have a mentor than control teachers. In the RCT sites, treatment teachers were more likely than control teachers to have a formally assigned mentor in both years of teaching (Exhibit 11).[13] Treatment teachers across all three sites met with their mentors more frequently and for more time than control teachers (Exhibits 11 and 12).

Exhibit 11. New Teacher Interactions with Mentors, RCT Combined Sample, 2014–16



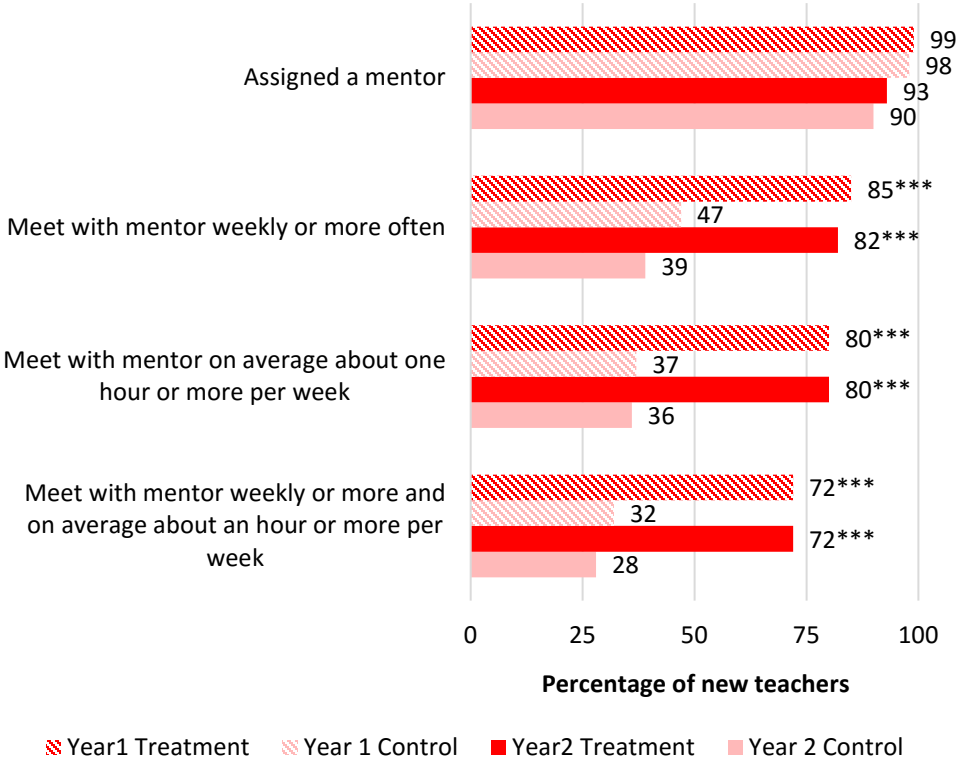
Source: NTC New Teacher Survey, spring 2014–2016.

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

Value labels are rounded to the nearest percentage point.

[13] Although all teachers in the treatment group were assigned mentors, a couple reasons might account for why only 95% of treatment teachers reported having a mentor. Self-report error is a potential risk in any survey; here, we assume that self-report error affects both treatment and control groups equally. Also, in a few rare cases reported through interviews, beginning teachers refused to meet regularly with their mentors and might have thus reported on the survey that they did not have a mentor.

Exhibit 12. New Teacher Interactions with Mentors, QED Site, 2013–16



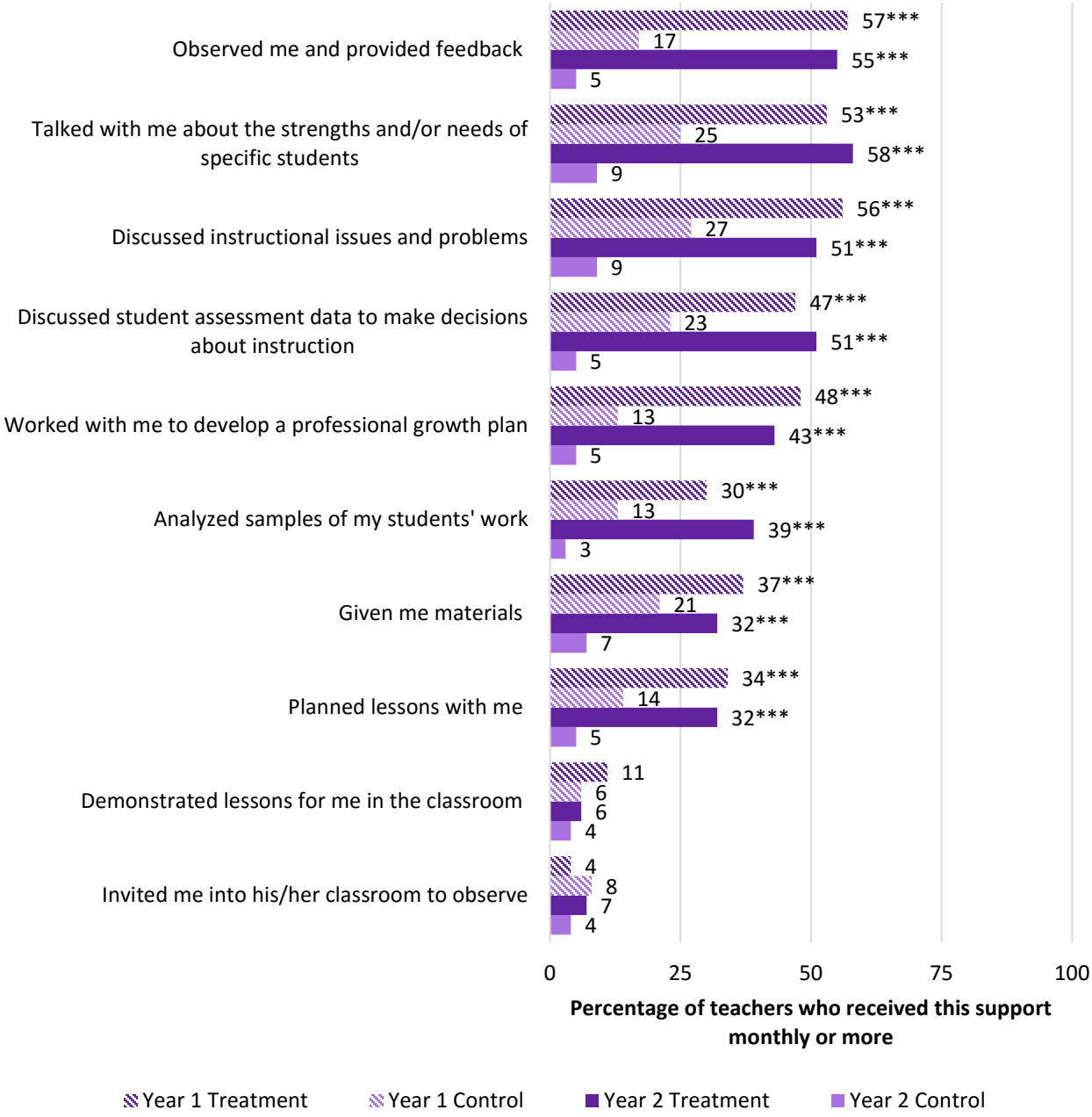
Source: NTC New Teacher Survey, spring 2013–2016.

*p < .05, **p < .01, ***p < .001.

Value labels are rounded to the nearest percentage point.

NTC mentors in all three sites and in both years worked with new teachers more consistently than non-NTC mentors in observing instruction and providing feedback, talking with teachers about the strengths or needs of specific students, discussing student assessment data to make decisions about instruction, and working with teachers to develop a professional growth plan (Exhibits 13 and 14). Interviews with treatment teachers corroborated that frequent and consistent mentoring mattered—teachers knew they had a knowledgeable and supportive colleague to rely on, who could observe and provide feedback regularly and get a feel for the classroom and respond to the teacher's needs.

Exhibit 13. Frequency of Mentoring Activities, RCT Combined Sample, 2014–16

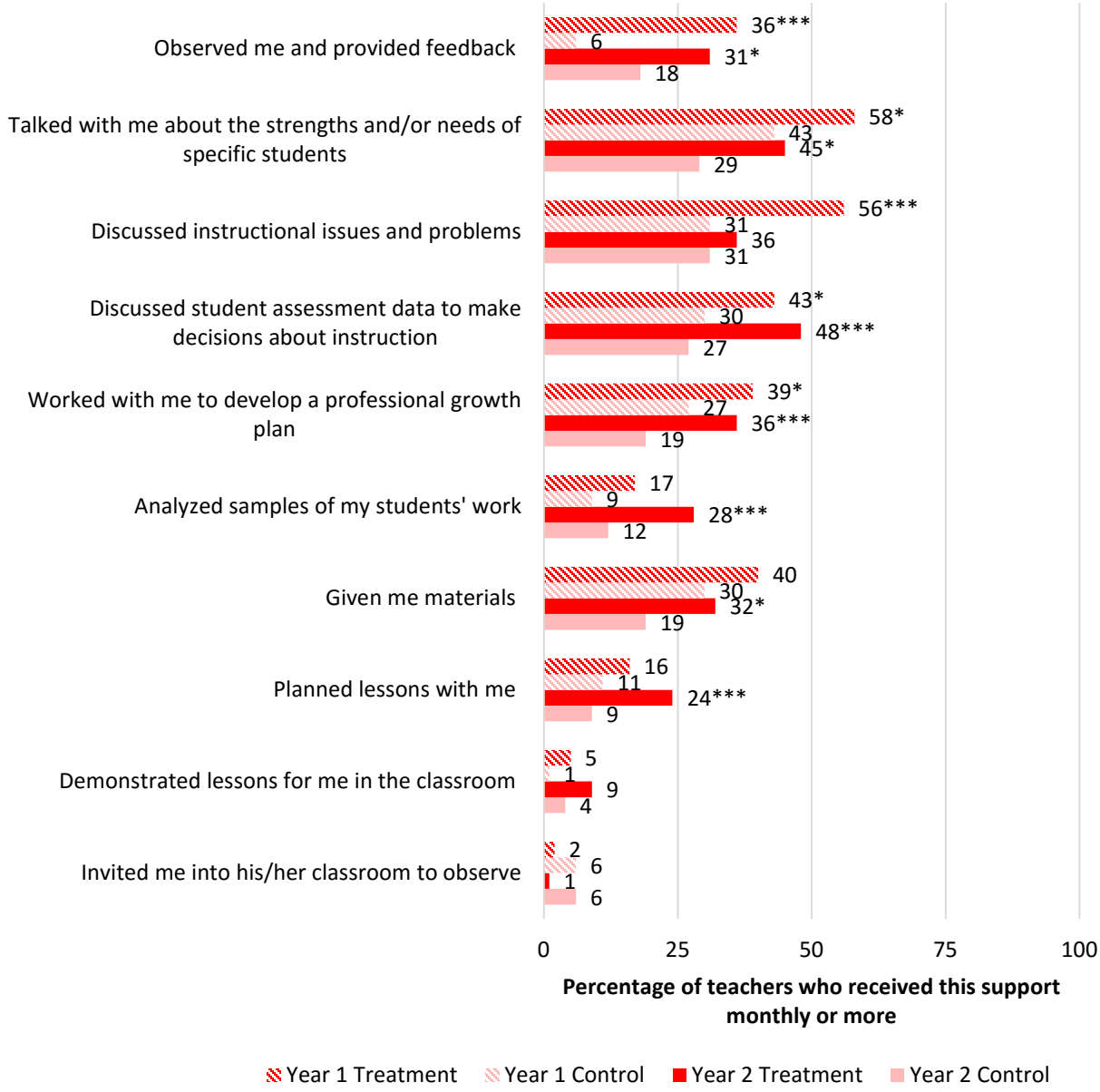


Source: NTC New Teacher Survey, spring 2014–2016.

*p < .05, **p < .01, ***p < .001.

Value labels are rounded to the nearest percentage point.

Exhibit 14. Frequency of Mentoring Activities, QED Site, 2013–16



Source: NTC New Teacher Survey, spring 2013–2016.

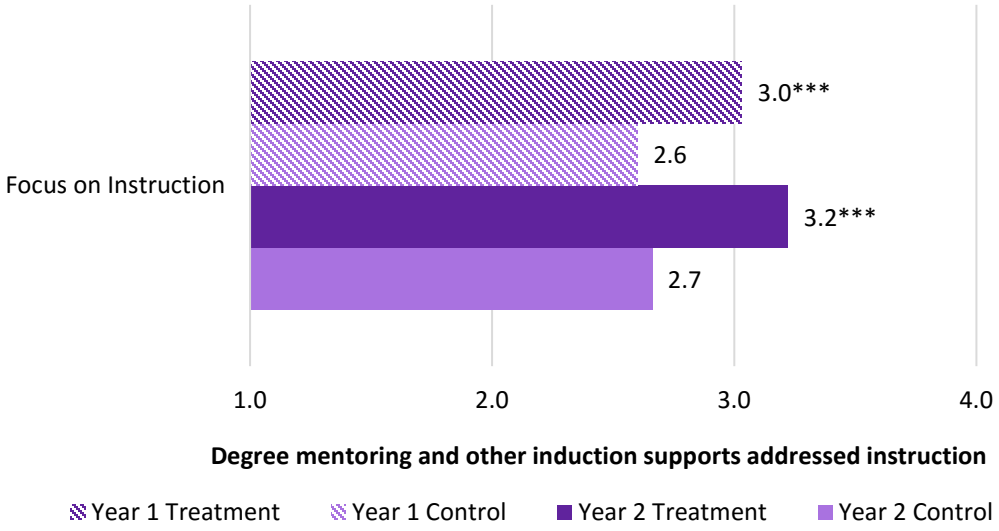
*p < .05, **p < .01, ***p < .001.

Value labels are rounded to the nearest percentage point.

The mentoring treatment teachers received focused more on instruction than that control teachers received. Treatment teachers in all three sites and in both years were more likely than control teachers to report that induction supports addressed instruction (Exhibits 15 and 16). In the RCT sites in particular, NTC teachers were more likely to report that mentoring and other induction supports focused on evaluating and reflecting on their teaching practice, adapting

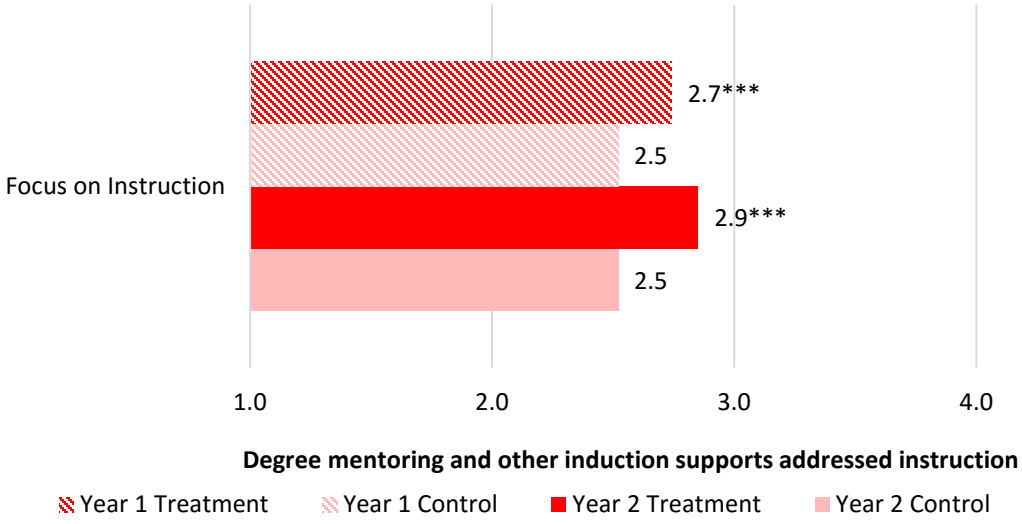
instruction to meet the needs of students at varying academic levels, instructional techniques appropriate to the grade level and subject matter they taught, and the use of assessment strategies in instruction (Exhibits 17 and 18).

Exhibit 15. Focus on Instruction in Mentoring and Other Induction Supports, RCT Combined Sample, 2014–16



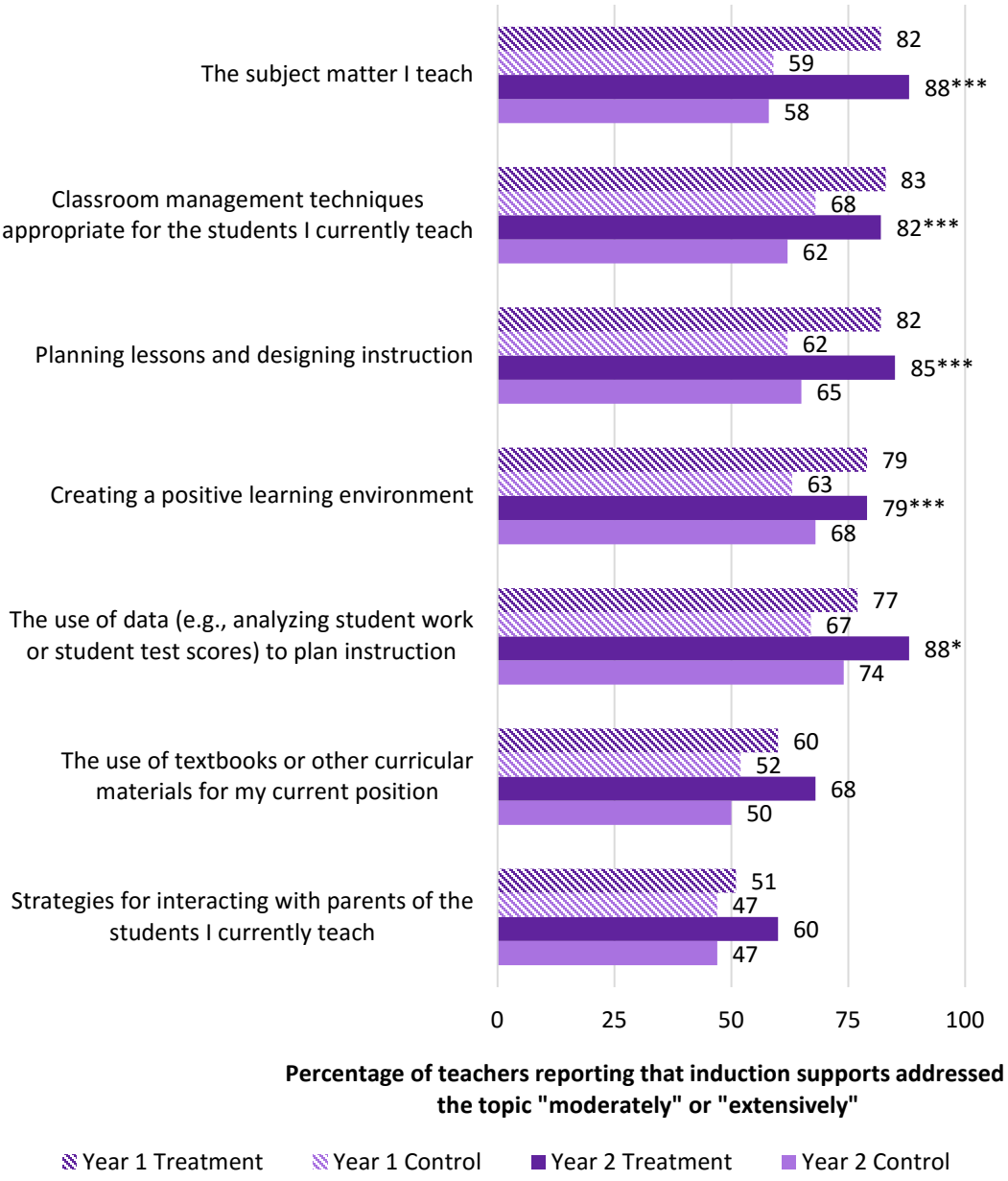
Source: NTC New Teacher Survey, spring 2014–2016.
*p < .05, **p < .01, ***p < .001.
Value labels are rounded to the nearest tenth.

Exhibit 16. Focus on Instruction in Mentoring and Other Induction Supports, QED Site, 2013–16



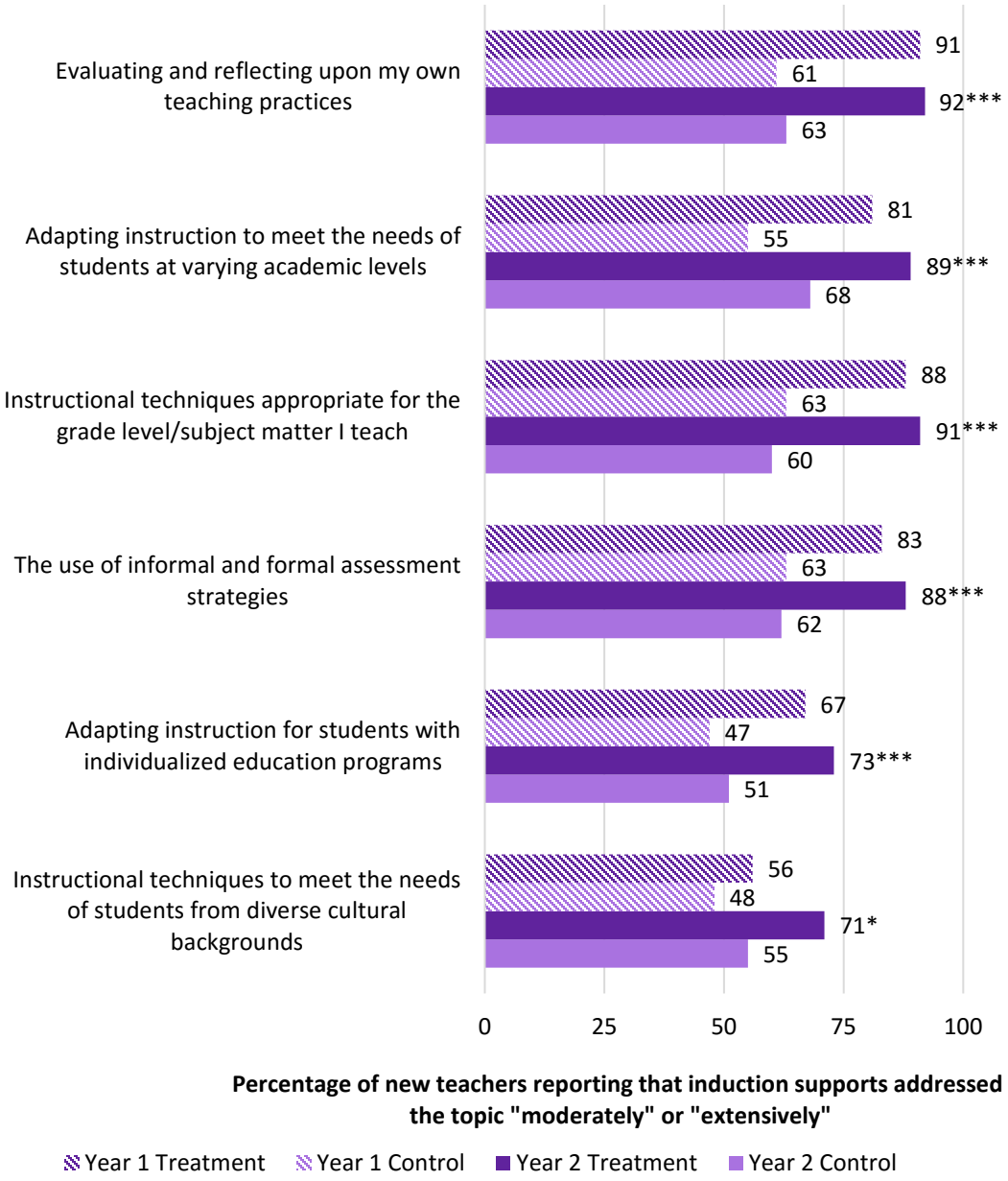
Source: NTC New Teacher Survey, spring 2013–2016.
*p < .05, **p < .01, ***p < .001.
Value labels are rounded to the nearest tenth.

Exhibit 17. Focus of Mentoring and Other Induction Supports, RCT Combined Sample, 2014–16



Source: NTC New Teacher Survey, spring 2014–2016.
*p < .05, **p < .01, ***p < .001.
Values labels are rounded to the nearest percentage point

Exhibit 17. Focus of Mentoring and Other Induction Supports, RCT Combined Sample, 2014–16 (concluded)

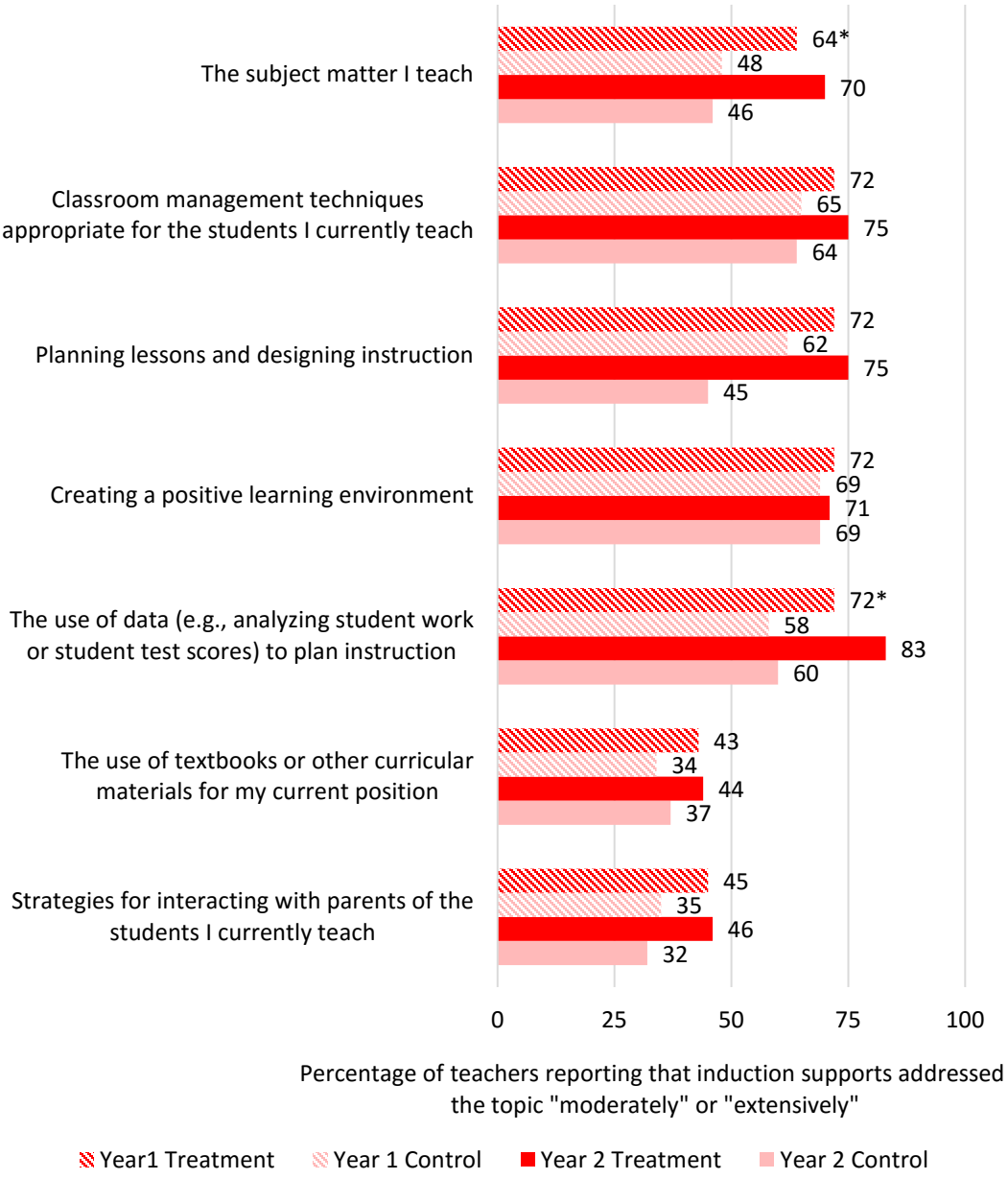


Source: NTC New Teacher Survey, spring 2014–2016.

*p < .05, **p < .01, ***p < .001.

Value labels are rounded to the nearest percentage point.

Exhibit 18. Focus of Mentoring and Other Induction Supports, QED Site, 2013–16

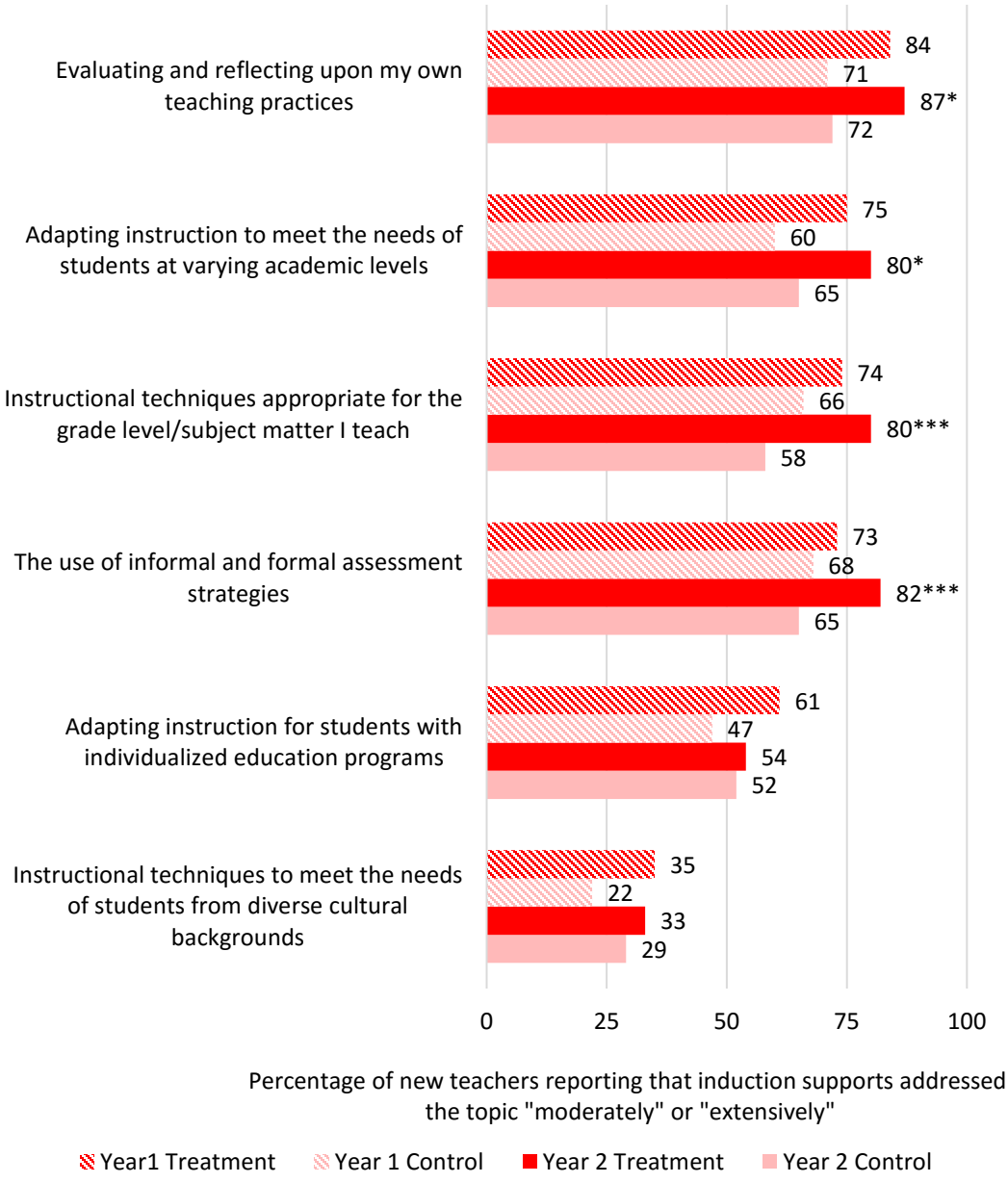


Source: NTC New Teacher Survey, spring 2013–2016.

*p < .05, **p < .01, ***p < .001.

Value labels are rounded to the nearest percentage point.

Exhibit 18. Focus of Mentoring and Other Induction Supports, QED Site, 2013–16 (concluded)



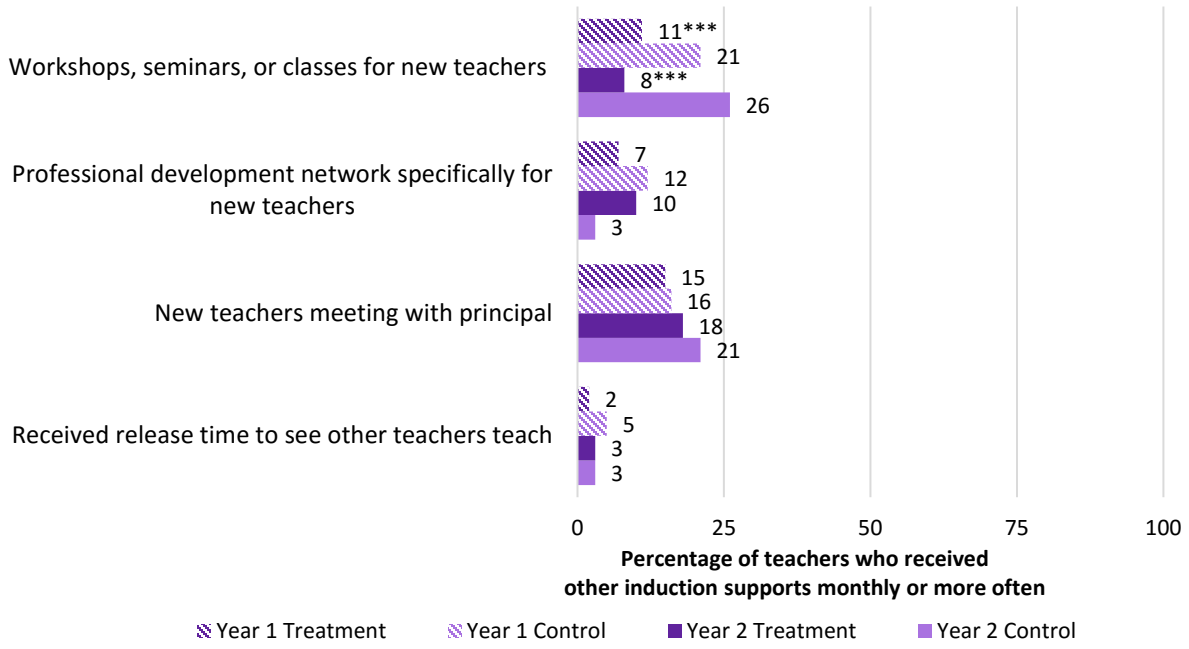
Source: NTC New Teacher Survey, spring 2013–2016.

*p < .05, **p < .01, ***p < .001.

Value labels are rounded to the nearest percentage point.

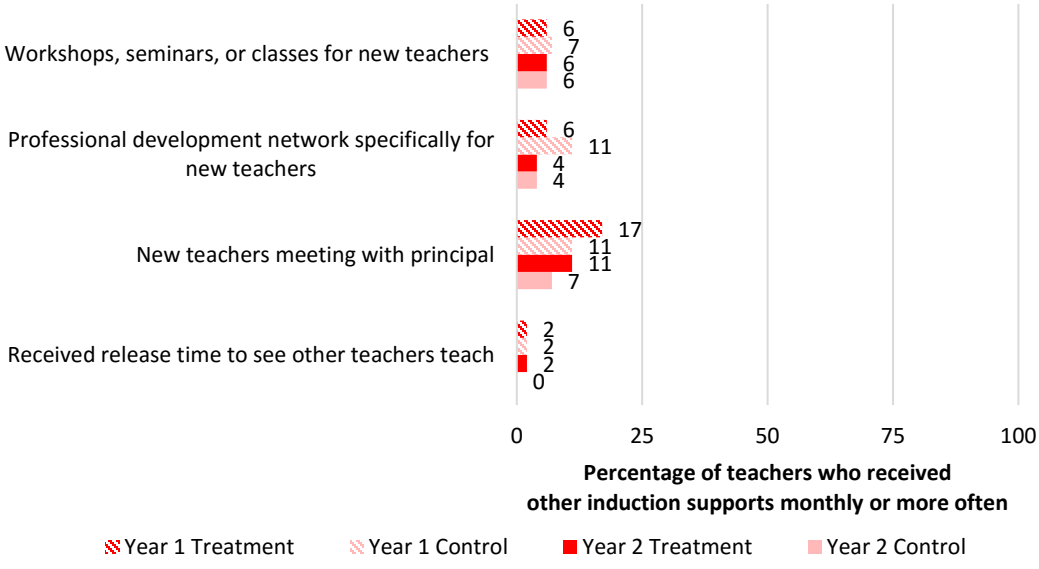
Access to induction supports other than mentors was limited for both treatment and control teachers. Overall, the percentages of teachers who received at least monthly induction supports other than mentoring were small (Exhibits 19 and 20), although in both years control teachers in the RCT districts were more likely than treatment teachers to attend workshops, seminars, or classes for new teachers (Exhibit 19).

Exhibit 19. Frequency of Other Induction Supports, RCT Combined Sample, 2014–16



Source: NTC New Teacher Survey, spring 2014–2016.
*p < .05, **p < .01, ***p < .001.
Value labels are rounded to the nearest percentage point.

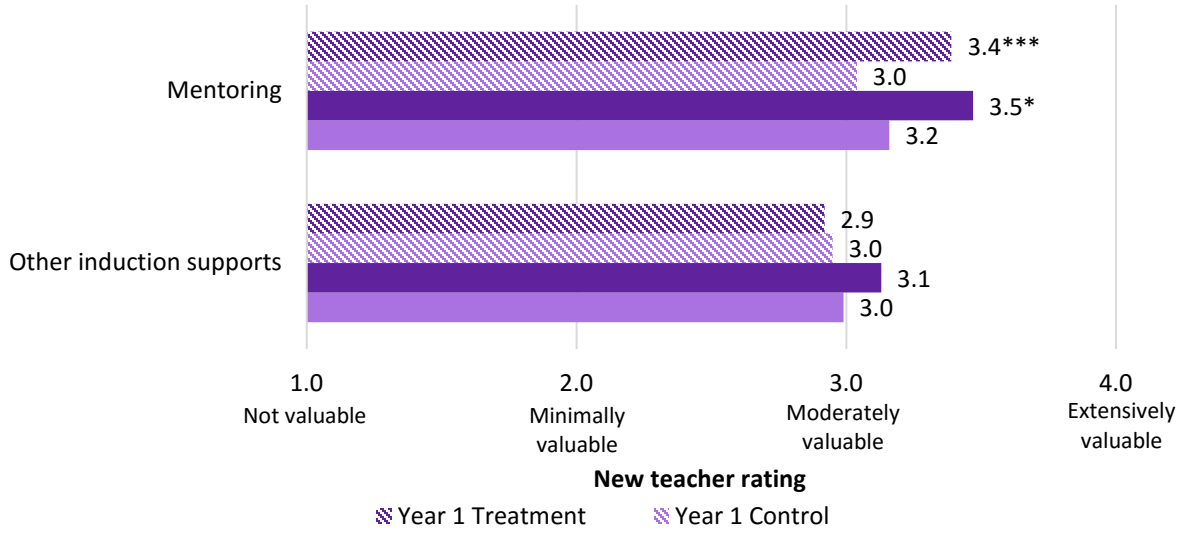
Exhibit 20. Frequency of Other Induction Supports, QED Site, 2013–16



Source: NTC New Teacher Survey, spring 2013–2016.
*p < .05, **p < .01, ***p < .001.
Value labels are rounded to the nearest percentage point.

Treatment teachers, compared with control teachers, rated the value of the mentoring activities they engaged in higher. Treatment teachers in the RCT sites in both years perceived more value in their mentoring activities than control teachers, as did QED site teachers in their second year of teaching (Exhibits 21 and 22).

Exhibit 21. Value of Mentoring Activities and Other Induction Supports, RCT Combined Sample, 2014–16

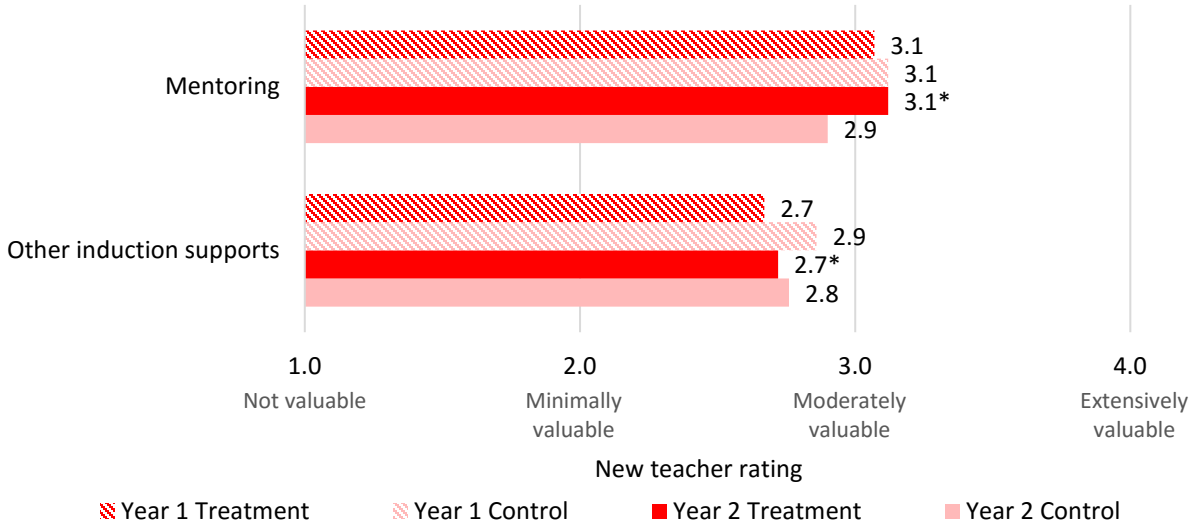


Source: NTC New Teacher Survey, spring 2014–2016.

*p < .05, **p < .01, ***p < .001.

Value labels are rounded to the nearest tenth.

Exhibit 22. Value of Mentoring Activities and Other Induction Supports, QED Site, 2013–16



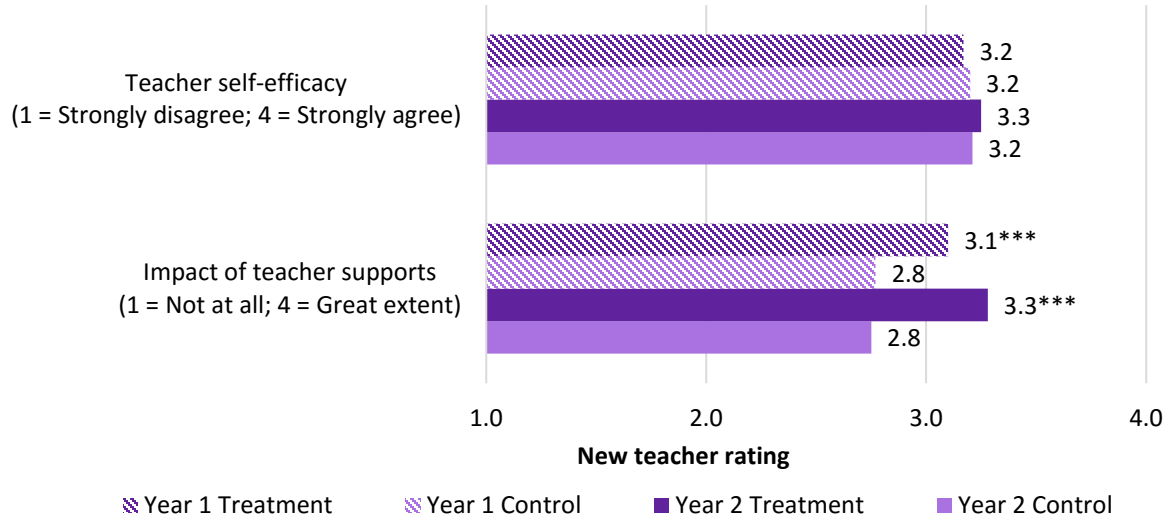
Source: NTC New Teacher Survey, spring 2013–2016.

*p < .05, **p < .01, ***p < .001.

Value labels are rounded to the nearest tenth.

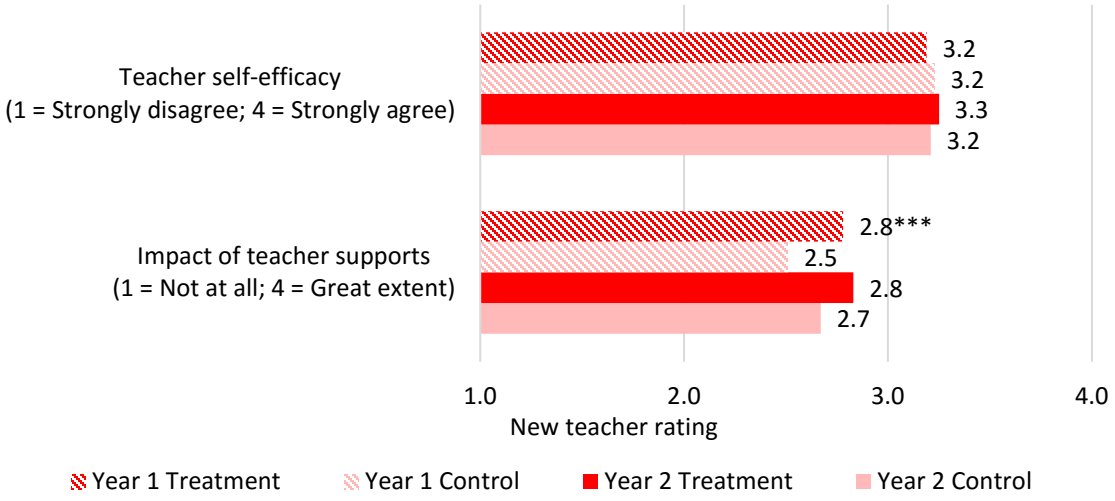
Further reflecting teachers' perception of the value of mentoring, treatment teachers in all sites were more likely than control teachers to report that induction supports helped them grow as teachers, even though treatment and control teachers had similarly positive assessments of their own efficacy in the classroom (Exhibits 23 and 24). Interviewed mentors and teachers elaborated that the 2-year model allowed mentors to use a gradual release approach in having teachers, over time, more independently use high-leverage protocols and tools such as analyzing student work and data to drive their own reflection and growth. Teachers built confidence and independence while continuing to receive support for 2 years.

Exhibit 23. New Teacher Ratings of Self-Efficacy and Impact of Supports, RCT Combined Sample, 2014–16



Source: NTC New Teacher Survey, spring 2014–2016.
*p < .05, **p < .01, ***p < .001.
Value labels are rounded to the nearest tenth.

Exhibit 24. New Teacher Ratings of Self-Efficacy and Impact of Supports, QED Site, 2013–16



Source: NTC New Teacher Survey, spring 2013–2016.
*p < .05, **p < .01, ***p < .001.
Value labels are rounded to the nearest tenth.

These results suggest that the significantly different levels of mentoring that treatment teachers and control teachers received on instructional priorities was consistent and persistent. Moreover, control teachers did not participate more than treatment teachers in other induction supports that might have compensated for the differences in their mentoring experiences. Thus, the overall high levels of implementation fidelity and treatment-control contrast together suggest that the participating sites provided a good testing ground for the impact of the NTC induction model on teacher and student outcomes.

NTC induction, as depicted in the logic model, is intended to improve beginning teachers' practice through its comprehensive support system as the path toward improved student learning. More robust support during those critical beginning years in the teaching profession is also intended to retain teachers at higher rates than districts would achieve if beginning teachers were left to struggle on their own. We present results for impacts on teacher practice and teacher retention for the RCT sites after 2 years of induction supports.[14]

Impacts on Teacher Practice

We measured teacher practice outcomes through structured classroom observations using the Framework for Teaching (Danielson, 2013), capturing dimensions of classroom management and culture and instructional quality on the same sample at baseline (at the beginning of the teachers' first year of teaching) and at the end of the teachers' second year of teaching.

Final Observation Analysis Sample

The analysis of classroom observations of teacher practice included treatment and control teachers who were randomly selected and observed in fall 2013 (baseline) and spring 2015 (after 2 years of teaching) for Cohort 1 and in fall 2014 (baseline) and spring 2016 (after 2 years of teaching) for Cohort 2. Teachers were eligible for the sample if they taught core subjects (mathematics, reading/English language arts, social studies, science, or self-contained elementary classrooms). We conducted all observations during instruction in the core subjects.[15]

Attrition

We measured attrition from the observation sample at the school level. Schools attrited from the sample when all teachers who were selected for observation in the school attrited, i.e., no teachers selected for observation were observed at both time periods. Exhibit 25 displays the number of treatment and control schools with teachers selected for observation in each cohort,[16] the number of schools with teachers observed at each time period, and the school-level attrition by condition in each district and overall. WWC standards for attrition take into account both overall attrition and the difference in attrition between treatment and control groups. In Cohort 1, overall school-level attrition was 36 percent, with differential attrition of 3 percentage points. In Cohort 2, overall attrition was 21 percent, with differential attrition of 6 percentage points. When combined,

[14] We were not able to include the QED site in the teacher outcomes. For teacher practice, we could not observe the comparison cohort of beginning teachers when they first started teaching because that period preceded the start of the grant. Given that veteran teachers (needed in the difference-in-differences design) do not serve as a sound comparison group for new teacher retention because they inherently have different retention patterns from new teachers, the teacher retention analysis from the QED site is a purely descriptive off-year comparison; therefore, we conducted only descriptive, not causal, analysis to inform NTC.

[15] Observers were trained and calibrated on the Framework for Teaching before each round of observations. After calibrations, observers achieved interrater reliability where over 90% of scores were within one point across all elements scored across multiple test videos.

[16] This number includes all schools with teachers selected, including those who declined to participate and the teachers who replaced them. In some cases, the teachers selected as replacements also declined to participate. Therefore, the total number of schools selected may have been larger in one district or condition than in the other, with the aim of obtaining a final sample that was balanced across treatment and control in each district.

the attrition for both cohorts was 28 percent, with differential attrition of 2 percentage points. These rates are all within the range of acceptable attrition.

After attrition, the number of schools remaining in the analysis sample was relatively low, as was the number of teachers in each of these schools, even when both cohorts and districts were combined. Lower sample size limited our ability to detect the effects of the NTC model on teacher practice, particularly if those effects were small or variability in practice among teachers was great. A second consequence of attrition, particularly when differential attrition occurs, is that the schools and teachers remaining in the sample may differ in both measurable and unmeasurable ways from those who attrited. Because the differential attrition was relatively low, this implication was not particularly problematic in the observation sample.

Exhibit 25. Overall School Observation Sample Selection and Attrition, Cohort 1, Cohort 2, and Combined, RCT Districts

	Cohort 1			Cohort 2			Combined		
	Treat-ment	Control	Over-all	Treat-ment	Control	Over-all	Treat-ment	Control	Over-all
Schools selected for observation (with replacement)	46	41	87	45	42	87	91	83	174
Observed at time 1	44	37	81	45	42	87	89	79	168
Stayed and were observed at time 2	**29**	**27**	**56**	**37**	**32**	**69**	**66**	**59**	**125**
Percentage attrited	37%	34%	36%	18%	24%	21%	27%	29%	28%
	Met standard			*Met standard*			*Met standard*		

Baseline Equivalence

To address the concern that teachers who remained in the analysis may be different from those who attrited, both on observable and unobservable factors, WWC requires that we show the baseline equivalence on the outcome measures of treatment and control teachers if attrition exceeds acceptable levels. This step was not necessary because the attrition levels met WWC standards. However, a full discussion of baseline equivalence is presented in Appendix D.

Measures of Teacher Practice

The teacher practice outcomes were eight measures on the Framework for Teaching (Danielson, 2013), four components each under Domain 2: Classroom Environment and Domain 3: Instruction. Trained observers scored each observed teacher on the 12 elements representing the four components under Domain 2 and the 15 elements representing four components under Domain 3, as shown in Exhibit 26.

The scores used in the analysis were factor variables combining the element-level scores into one variable representing each component; each component was based on two to four elements (Exhibit 26).[17] Each factor variable was continuous, had a mean of 0 and a standard deviation of 1, and the majority of teachers scored in the range from -2 to 2. A score of zero on each component therefore is equivalent to being at the average score for teachers in this sample at baseline. A change in these variables of 1.0 is a change of 1 standard deviation, which is roughly equivalent to 0.5 or 0.6 point on the original 1 to 4 scale on the Framework for Teaching.

[17] The factor variable reflects the structure of the correlations between the elements. It is similar to a weighted average of the elements, where the weights include the strength of the relationship between the elements as well as teachers' scores on the elements.

Exhibit 26. Framework for Teaching Domains, Components, and Elements Observed

Domain 2. The Classroom Environment	
Component (Factor Variable)	**Elements**
Creating an Environment of Respect and Rapport	Teacher interactions with students Student interactions with other students
Establishing a Culture for Learning	Importance of the content and of learning Expectations for learning and achievement
Managing Classroom Procedures	Management of instructional groups Management of transitions Management of materials and supplies Performance of classroom routines
Managing Student Behavior	Expectations Monitoring of student behavior Response to student misbehavior
Domain 3: Instruction	
Component (Factor Variable)	**Elements**
Communicating with Students	Expectations for learning Directions for activities Explanations of content Use of oral and written language
Using Questioning and Discussion Techniques	Quality of questions/prompts Discussion techniques Student participation
Engaging Students in Learning	Activities and assignments Grouping of students Instructional materials and resources Structure and pacing
Using Assessment in Instruction	Assessment criteria Monitoring of student learning Feedback to students Student self-assessment and monitoring of progress

Source: Excerpted from the Framework for Teaching (Danielson, 2013).

Controls

The teacher practice impact estimates were derived from a two-level hierarchical model with teacher and school levels that controlled for school- and teacher-level variables, the baseline observation measure for each teacher observed, and the blocking variables used in the random selection and assignment of schools. Using these blocking variables ensured representation of different types of schools by geography (in CPS), Teacher Incentive Fund status (in BCPS), and grade levels served (in both districts).

Results

Examining the teacher practice outcomes for the combined RCT districts, we found no statistically significant differences between observed treatment and control teachers on any of the measured components in either cohort or when the two cohorts were combined. Exhibit 27displays the impact estimates for each component for the combined RCT districts.[18] The small sample size in

[18] Because the relationship between the control variables and the outcomes differed by cohort, as did the structure of the variation at the school and teacher levels, the estimates found in the models by cohort do not average to the estimates found in the analysis including both cohorts.

the end impeded our ability to find impacts. Post hoc power tests indicated that the final sample size had minimum detectable effect sizes of 0.40 to 0.46 across the eight teacher practice outcomes. Full model tables are in Appendix D.

Exhibit 27. Impact of the NTC Model on Teacher Practice Outcomes, Combined RCT Sample

		Cohort 1	Cohort 2	Both Cohorts
Creating an Environment of Respect and Rapport	Estimate	0.16	-0.10	0.04
	SE	(0.21)	(0.25)	(0.16)
	N teachers	71	88	159
	N schools	56	69	108
Establishing a Culture for Learning	Estimate	-0.10	-0.14	-0.24
	SE	(0.23)	(0.34)	(0.20)
	N teachers	71	88	159
	N schools	56	69	108
Managing Classroom Procedures[a]	Estimate	0.23	0.07	0.13
	SE	(0.25)	(0.23)	(0.16)
	N teachers	59	80	139
	N schools	51	68	102
Managing Student Behavior[b]	Estimate	0.34	0.18	0.28
	SE	(0.24)	(0.24)	(0.18)
	N teachers	71	88	159
	N schools	56	69	108
Communicating with Students	Estimate	-0.06	0.08	0.01
	SE	(0.24)	(0.27)	(0.19)
	N teachers	70	87	159
	N schools	56	69	108
Using Questioning and Discussion Techniques	Estimate	-0.04	0.28	0.21
	SE	(0.23)	(0.28)	(0.18)
	N teachers	69	88	157
	N schools	55	69	107
Engaging Students in Learning	Estimate	-0.11	0.43	0.15
	SE	(0.24)	(0.28)	(0.17)
	N teachers	70	87	157
	N schools	56	68	108
Using Assessment in Instruction	Estimate	-0.21	0.18	0.06
	SE	(0.22)	(0.30)	(0.18)
	N teachers	70	88	158
	N schools	55	69	107

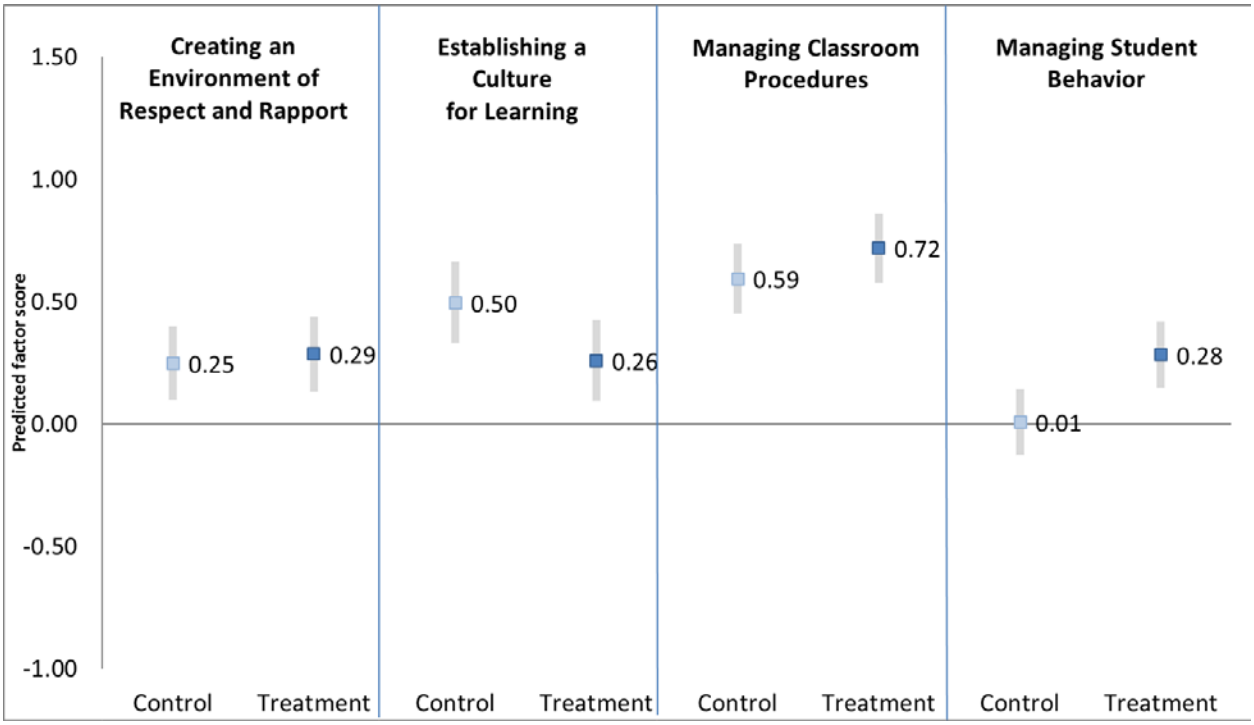
*$p < .05$, ** $p < .01$, *** $p < .001$.

[a] This variable excludes the elements that had many blanks because *Management of Instructional Groups* and *Performance of Classroom Routines* were not observed during the observation period for a number of teachers.

[b] This variable excludes the elements that had many blanks because *Response to Student Misbehavior* was not observed during the observation period for a number of teachers.

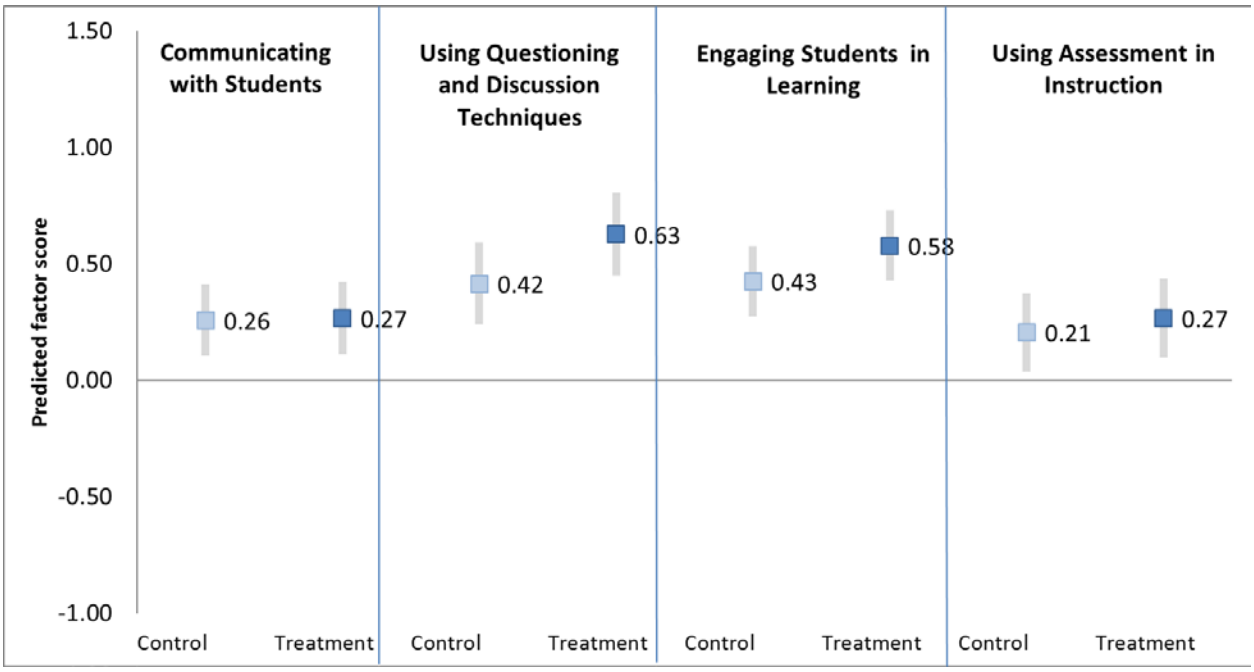
Exhibit 28 displays the results of the impact analysis combining both districts and cohorts. The estimates for treatment and control account for the baseline observation score for teachers, as well as teacher and school demographics. As discussed, factor scores were created at baseline to have a mean of 0 and a standard deviation of 1. Therefore, the positive estimates for both treatment and control indicate that on average the sample of teachers at the end of 2 years scored higher on the Framework for Teaching measures than the sample at baseline. However, the amount of growth on these measures was not significantly different between teachers in treatment and control groups. Additionally, these estimates confound the impact of teacher growth and the impact of attrition; the average score may be higher at the end of 2 years because of attrition of lower performing teachers from the sample. On three factors, *Creating an Environment of Respect and Rapport*, *Communicating with Students*, and *Using Questioning and Discussion Techniques*, the baseline scores of teachers who attrited were significantly lower than the baseline scores of teachers who remained in the sample. Such a pattern suggests that weaker teachers may have attrited from the sample overall, consistent with the pattern of higher mean scores compared with baseline across all teacher practice measures and for both treatment and control groups.

Exhibit 28. Model-Implied Means on Teacher Practice Outcomes for Treatment and Control Groups Overall



Note: The blue boxes depict the estimated treatment and control means, and the grey lines depict the standard error around those means. The standard error is a measure of confidence in the estimate of the mean, and when the grey lines of treatment and control overlap, we cannot say with confidence that the means are significantly different between the two groups.

Exhibit 28. Model-Implied Means on Teacher Practice Outcomes for Treatment and Control Groups Overall (concluded)



Note: The blue boxes depict the estimated treatment and control means, and the grey lines depict the standard error around those means. The standard error is a measure of confidence in the estimate of the mean, and when the grey lines of treatment and control overlap, we cannot say with confidence that the means are significantly different between the two groups.

Impact on Teacher Retention

Using district administrative data, SRI assessed the impacts of the NTC induction model on teachers' retention into their third year of teaching (a 2-year retention rate) in the RCT districts.

Analysis Sample

Human resources data were complete for all study teachers in Cohort 1 and Cohort 2 in treatment and control schools in the RCT sites. Therefore, no attrition occurred in the sample for the teacher retention analysis.

Outcome Measure

Because summers are a natural time for teachers to change jobs, we counted teachers as "retained" if they were still employed by the district at the beginning of their third year, i.e., fall 2015 for Cohort 1 and fall 2016 for Cohort 2.

Results

We estimated the impact of the NTC induction model on teacher retention using a two-level hierarchical model with the same controls as those in the teacher practice models. Across both cohorts and districts, 79 percent of treatment teachers and 78 percent of control teachers were retained; the difference was not statistically significant. The retention rates for both treatment and control teachers are lower than the rate found for a national sample of teachers that began teaching in 2007–08, 85 percent of whom remained in the profession 3 years later (Gray & Taie, 2015). Local factors, such as one participating district being commonly thought of as a teacher training ground for suburban districts, might be stronger than the any mitigation offered by stronger induction supports. See Appendix E for full teacher retention models.

The NTC induction model aims not only to provide novice teachers with instructional skills, but also to cultivate in them the reflective skills to critically analyze their own practice and use data as evidence of instruction that leads to deep student learning. This chapter reports on the studies of impacts on student achievement, the RCTs in two districts and the QED in the third site.

Randomized Controlled Trials of the Impact on Student Achievement

Through the RCTs, SRI examined the impact of the NTC induction model on English language arts and mathematics achievement in grades 4 through 8 in teachers' second year of participating in NTC's 2-year induction support. Special education teachers who taught reading and/or mathematics and who could be linked to students in district data sets were included in the analysis, along with regular education teachers.

Attrition

Overall attrition at the school level for the combined sample across the RCT districts was 1 percent for mathematics and 10 percent for ELA. Both the overall and differential attrition levels were within acceptable thresholds under WWC (see Appendix F).

Baseline Equivalence in RCT Districts

We examined baseline equivalence in student achievement scores between treatment and control schools. Baseline equivalence ensures that any differences in the outcomes between treatment and control groups are due to the treatment and not to systematic differences between the groups that were present before the intervention. A baseline difference of less than 0.05 standard deviation is considered equivalent. A baseline difference between 0.05 and 0.25 can be considered equivalent if prior achievement is included in the model. We included prior achievement in the models for all student achievement analyses to obtain better precision for the estimated impact.

For the second-year impact, with the two RCT sites and both cohorts combined, the difference in baseline achievement scores between students with treatment teachers and students with control teachers was 0.01 standard deviation in mathematics and 0.10 standard deviation in ELA. As the difference in mathematics was below 0.05 standard deviation, this analysis achieved baseline equivalence. The difference in ELA was between 0.05 standard deviation and 0.25 standard deviation; therefore, it achieved baseline equivalence with prior achievement included in the model.

Student Achievement Measures

We used scale scores from state assessments of ELA and mathematics for grades 4 through 8 as measures of student achievement.[19] In BCPS, we used scale scores from the 2014–15 and 2015–16 Florida Standards Assessment (FSA) (for Cohort 1 and Cohort 2 teachers in their second year of teaching, respectively), and similarly in CPS we used scale scores from the 2014–15 and 2015–16 Measures of Academic Progress (MAP).[20] To combine test results across grade levels and also

[19] Students in third grade take state assessments in Florida, Illinois, and Iowa. The third-grade scores serve as the measure of prior achievement for fourth-grade students. As the lowest tested grade, however, third-grade students do not have a measure of prior achievement and could not be included in the analysis. Fourth grade was the lowest grade that we could include in the sample.

[20] Because the state of Illinois changed assessments, CPS administered the MAP in 2014–15 and 2015–16 to bridge the two time periods under different assessments. CPS needed continuity in student achievement measurements for the district's teacher evaluation system.

across districts, we standardized each scale score at each grade level using a common metric, the *z* score, which has a mean of 0 and a standard deviation of 1.[21]

We compared the test scores of students of treatment teachers with those of control teachers, controlling for the prior achievement of each student (as measured by their test scores in 2013–14 or 2014–15), student background, teacher background, school characteristics, and the blocking variables used in the random assignment of schools into treatment and control groups. These blocking variables ensured that the sample represented different types of schools by geography (in CPS), Teacher Incentive Fund status (in BCPS), and grade levels served (in both sites). The hierarchical models that we used to estimate the impact of NTC induction accounted for the nesting of students within classrooms and of teachers within schools. The blocking variables accounted for both sampling design and district-level nesting because schools in each site were blocked on a different set of variables. Because this analysis combined data collected in two different years (2014–15 for Cohort 1 and 2015–16 for Cohort 2), a centered year variable was also included to account for any historical changes in test scores between the two years. Finally, we included interactions between the district and cohort indicators and all background characteristics to account for the different relationships between background characteristics and student outcomes by district and cohort.

Second-Year Impact, Combined RCT Sites

Exhibit 29 shows the difference between the adjusted mean test scores among treatment teachers' students and control teachers' students at the end of the second year of teaching, controlling for prior achievement, student characteristics, teacher characteristics, and school characteristics. This difference, measured in standard deviations of the underlying distribution of student scale scores across both sites, represents the 1-year impact of NTC induction on the student achievement of teachers in their second year of induction support.

In ELA, the average student achievement of teachers in the second year who participated in NTC induction for 2 years was approximately 0.05, compared with -0.04 for students of control teachers. This difference equals an effect size of 0.09 standard deviation ($p < .05$)—equivalent to moving from the 48th to the 52nd percentile. On broad-scope standardized tests of reading like the FSA and the MAP, an effect size of 0.09 is equivalent to an approximately 23 to 39 percent greater annual gain than otherwise expected for students in grades 4 through 8 and represents the equivalent of approximately 2 to 3.5 additional months of learning, depending on the student's grade level (Lipsey et al., 2012).

The NTC induction model also showed significant and positive impacts on mathematics achievement in grades 4 through 8. Students in grades 4 through 8 of teachers in their second year who participated in NTC induction for 2 years scored 0.15 standard deviation ($p < .01$) higher on average than students of control teachers. These impacts are equivalent to moving from the 46th to the 52nd percentile. On broad-scope standardized tests like the FSA and the MAP, an effect size of 0.15 is equivalent to an approximately 27–50 percent greater annual gain than otherwise expected for students in grades 4 through 8 and represents the equivalent of approximately 2.4 to 4.5 additional months of learning, depending on the student's grade level.

[21] To calculate *z* scores, we first computed the mean and standard deviation of scale scores separately for ELA and for mathematics in each site and at each grade level, based on the full set of scores that we received for students of treatment and control teachers. We converted scale scores to *z* scores by taking the scale score in ELA or mathematics, subtracting the overall sample mean for that subject, and dividing by the pooled standard deviation for students of treatment and control teachers. A *z* score of 0 means that the student scored at the mean for his or her grade level in the study schools in his or her district in ELA or in mathematics. A *z* score of 1 means that the student scored 1 full standard deviation above the mean, and a *z* score of -1 means that the student scored 1 full standard deviation below the mean.

Exhibit 29. Second-Year Impact on Student Achievement, Combined RCT Sites

Subject	Adjusted Mean Test Scores		Difference	Sample Sizes		
	Treatment	Control	(effect size)	Students	Teachers	Schools
ELA	0.05	-0.04	**0.09***	6,147	149	99
Mathematics	0.06	-0.09	**0.15****	4,972	129	86

Note: The effect on student achievement is a 1-year effect as the districts provided current and prior achievement data annually but did not consistently provide identifiers to link students across the data sets given to researchers each year. The 1-year impact after 2 years of mentoring includes achievement in 2014–15 for Cohort 1 teachers and 2015–16 for Cohort 2 teachers.

Adjusted mean test scores are in standard deviation units.

* $p < .05$, ** $p < .01$.

We further tested the robustness of these findings by running several sensitivity analyses, including removing late joiners from the analytic sample (Appendix F), which did not change the student achievement results substantively. Appendix G provides other exploratory analyses on the RCT districts, including examining whether the results differ for elementary versus middle school students, for students taking ELA or mathematics with more than one teacher, or by school characteristics.

Quasi-experimental Study of the Impact on Student Achievement

In the quasi-experimental study, SRI used a differences-in-differences approach to estimate the impact of participating in the 2-year NTC induction program in one site. The study compared the difference in the 2014–15 achievement of students of Cohort 1 beginning teachers receiving NTC induction support from 2013–14 to 2014–15 and the 2013–14 student achievement of a prior cohort of comparison beginning teachers who started teaching in 2012–13 and did not receive NTC induction support with the difference in the student achievement of veteran teachers in the same years.[22]

Outcome Measures and Timing

The outcomes used in the QED were Iowa Assessment scores in ELA and mathematics in grades 4 through 8.[23] To combine test results across grade levels, we standardized each scale score at each grade level using the *z* score.[24] The distribution of standardized outcome scores for each of the four groups of teachers included in each of the four impact analyses are presented in Appendix H.

The state testing schedule posed an unexpected constraint on the analysis. GWAEA districts varied in when they chose to administer the state test in ELA and mathematics (fall, winter, or spring). The time period of the achievement data therefore did not align perfectly with the treatment period. For example, for districts testing in the fall, the baseline measure was taken at the

[22] From communication between NEi3 and SRI research team, this difference-in-differences approach has better validity and is more likely to meet the WWC standards when treatment and comparison groups are no more than 1 year apart, although WWC changed the standards for difference-in-differences approaches in 2014, after we had designed the impact study for GWAEA and the site had begun serving new teachers. SRI conducted student outcomes analysis for GWAEA teachers in the second cohort to inform NTC about its program. The results are in Appendix H and are not intended for WWC review.

[23] Students in third grade take state assessments in Iowa. The third-grade scores serve as the measure of prior achievement for fourth-grade students. As the lowest tested grade, however, third-grade students do not have a measure of prior achievement and could not be included in the analysis. Fourth grade was the lowest grade that we could include in the sample.

[24] To calculate *z* scores, we computed the mean and standard deviation of scale scores separately for reading and for mathematics at each grade level, based on the whole sample of students in GWAEA for a given year. We converted scale scores to *z* scores by taking the scale score in reading or mathematics, subtracting the overall sample mean for that subject, and dividing by the overall sample standard deviation. A *z* score of 0 means that the student scored at the mean for his or her grade level among the sample of students included in the study. A *z* score of 1 means that the student scored one full standard deviation above the mean.

beginning or slightly after treatment teachers started receiving supports, and the first-year outcome measure was taken at the beginning of the following school year, with whatever effect summer loss or transition to a new teacher or a new school level might have had on the students' test scores. If the testing window did not vary between the comparison cohort and the treatment cohort, we could assume that any effects of the testing window would equally affect the comparison and treatment teachers. We included indicators of the testing schedule in the analytic model to adjust for its effect on the outcome analysis.

Students in most participating districts were tested in the spring of each year, while a few districts tested in the fall or winter. Some districts switched testing from fall in one year to spring in the next during the evaluation period.[25] As an example, Exhibit 30 details the timing of prior achievement and second-year outcomes for Cohort 1 and comparison teachers in their second year of teaching. The fall to spring testing scenario was included in the Cohort 1 Year 2 impact analysis because the comparison cohort had a similar testing window. In all analyses, there was fall to fall, spring to spring, and winter to winter testing in both groups; we therefore included these testing patterns in all analyses.

Exhibit 30. Timing of Prior Achievement and Outcome Scores, QED Site

	District Testing Window	Baseline Achievement	Year 2 Outcome
Comparison teachers (Teachers beginning teaching in 2012–13)	Winter	Winter 2013	Winter 2014
	Spring	Spring 2013	Spring 2014
	Fall	Fall 2013	Fall 2014
	Fall to spring	Fall 2012	Spring 2014
Cohort 1 teachers (Teachers beginning teaching in 2013–14)	Winter	Winter 2014	Winter 2015
	Spring	Spring 2014	Spring 2015
	Fall	Fall 2014	Fall 2015
	Fall to spring	Fall 2013	Spring 2015

Analysis Sample

The student achievement analysis sample for Cohort 1 included all NTC and comparison new teachers who taught reading and/or mathematics in grades 4 through 8, together with their corresponding comparison veteran teachers (with 3 or more years of experience) who taught the same grade levels as the NTC/comparison new teachers in the same school. As described, all included students must have had aligned pre- and post-test scores. Special education teachers who taught reading and/or mathematics and who could be linked to students in district data sets were included in the analysis, along with regular education teachers.

The small number of new teachers in tested grades and subjects who had students with aligned pre- and post-test scores limited the student achievement analysis. There were more than 100 new teachers in the Cohort 1 and comparison groups. Of those, only a fraction—much lower than we expected—taught reading or mathematics in grades 4 through 8 in each year. The number further dropped and was somewhat uneven across years after we excluded teachers who had no students with aligned pre- and post-test scores.[26] Exhibit 31 displays the numbers of teachers and students included in reading and mathematics analyses for Cohort 1.

[25] A few schools that did not test at all during a given year of study were excluded from the analysis.

[26] Cohort 1 Year 2 impact analysis had fall to spring testing in both treatment and comparison groups, while Cohort 2 Year 2 analysis did not. Therefore, although Cohort 1 and Cohort 2 analyses were supposed to use the same new teacher comparison group, there was a larger comparison group sample for Cohort 1 than for Cohort 2 in the Year 2 impact analysis because for comparability, comparison teachers with fall to spring testing were dropped in the Cohort 2 Year 2 analysis

Exhibit 31. Numbers of Schools, Teachers, and Students Included in Cohort 1 Year 2 Achievement Analyses, QED Site

	Comparison					Treatment				
		New Teachers		Veteran Teachers			New Teachers		Veteran Teachers	
	No. of Schools	No. of Teachers	No. of Students	No. of Teachers	No. of Students	No. of Schools	No. of Teachers	No. of Students	No. of Teachers	No. of Students
ELA	8	8	194	35	1305	15	19	340	59	1189
Math	6	7	299	26	1140	17	23	533	68	1822

Baseline Equivalence

We examined baseline equivalence among students of the four groups of teachers for each of the difference-in-differences analyses. The top two panels of Exhibit 32 provide summary information on baseline student reading and mathematics standardized scores for the Cohort 1 Year 2 impact analysis. When the differences in baseline student achievement scores among these four groups of teachers were all under 0.25 standard deviation, the subsequent analysis achieved baseline equivalence once prior achievement was included in the models. However, for the Cohort 1 Year 2 mathematics analysis, there were differences in baseline scores larger than the 0.25 standard deviation threshold among students of the four groups of teachers. We took an additional precaution in accounting for the differences between these groups of teachers. For each analysis where there was a greater than 0.25 standard deviation difference in prior achievement, we created a propensity score weight by predicting the likelihood of being a student of a Cohort 1 teacher based on the prior achievement variable. Then we used this propensity score as a weight on students in the other three groups. This weighting approach added an extra correction for imbalances in baseline student achievement among the four groups of teachers. The bottom panels of Exhibit 32 present the descriptive information of the post-weighting baseline scores of the four groups of students. After weighting there was no differential baseline score larger than 0.25 between any groups for any analysis. This weight was therefore included in the subsequent impact analysis of the outcome score.

Exhibit 32. Cohort 1 Year 2 Baseline Student Test Scores, by Groups of Teachers, QED Site

	Cohort 1 Teachers (Year 2 treated, 2014–15)	Comparison Teachers (Year 2, 2013–14)	Comparison Veteran Teachers for Cohort 1 Teachers (2014–15)	Comparison Veteran Teachers for Comparison Teachers (2013–14)
ELA				
Mean	-0.18	0.04	-0.16	0.01
SD	1.02	0.92	1.04	1.03
N students	340	194	1189	1305
Mathematics				
Mean	-0.26	0.09	-0.01	-0.09
SD	0.98	0.99	1.01	0.95
N students	533	299	1822	1140
Mathematics (weighted)				
Mean	-0.26	-0.11	-0.22	-0.27
SD	0.98	1.02	1.00	0.93
N students	533	299	1822	1140

Second-Year Impact, QED Site

For the Cohort 1 Year 2 impact analysis in the QED site, we compared the test scores of students of Cohort 1 teachers in Year 2 with those of comparison teachers in Year 2, adjusting for the difference with their corresponding comparison veteran teachers, and controlling for the prior achievement of each student, student background, teacher background, and school characteristics. The hierarchical models that we used to estimate the impact of NTC induction accounted for the nesting of students within teachers, and teachers within schools. (Because most districts included in the analysis had only one school in the study, a district level in the hierarchical model was not possible.)

Exhibit 33 shows the difference-in-differences estimate for Cohort 1 teachers in their second year of NTC induction, together with sample sizes at the student, teacher, and school levels. This difference-in-differences estimate, measured in standard deviations of the underlying distribution of student scale scores, represents the impact of NTC induction on student achievement. The impact estimate for Cohort 1 teachers in their second year of induction support is not statistically significant, suggesting no detected NTC impact on Cohort 1 teachers in the QED site.

Given the very small number of new teachers in Cohort 1 with tested students and the varying testing schedules, only a small sample of NTC and comparison new teachers could be included in each of the analyses. Such a small number of NTC and comparison new teachers may not be representative of all new teachers in the QED site over the years, and the small sample itself prohibits a reliable estimation of the NTC impact. In view of these concerns with the small sample and its representativeness, this analysis should be interpreted with caution, and the results are not conclusive—we do not know whether the NTC induction model had an impact on student achievement in the QED site.

Exhibit 33. Estimated Impact on Student Achievement for Cohort 1 Year 2, QED Site

	Estimate	*SE*	*p*-value	No. of Schools	No. of Teachers	No. of Students
ELA	-0.078	0.065	0.258	23	121	3028
Mathematics	0.011	0.042	0.752	22	125	3794

CONCLUSIONS AND IMPLICATIONS

NTC implemented its induction model in three sites with high fidelity, and the model had a strong positive influence on teachers' induction experiences. Differences in the induction received by teachers in the treatment schools and control schools were notable. Relative to control teachers, NTC teachers met with their mentors more often, worked on tasks more directly related to instruction, placed a higher value on mentoring activities, and were more likely to credit their induction experience with contributing to the development of their knowledge and skills as teachers. Interviewed teachers elaborated on the value of the NTC mentors in helping them learn to manage their classrooms, plan thoughtfully, differentiate instruction, reflect on their practice, and gradually become more confident in their teaching.

The impacts on teacher and student outcomes were more mixed. From the RCTs, we detected no differences between treatment and control teachers in measures of classroom environment and instruction domains using the Framework for Teaching (Danielson, 2013). The lack of impact was likely due to attrition and small sample size. In addition, it is possible that the measures of teacher practice were not fine-grained enough to capture the nature of NTC effects on instruction. Retention rates into the third year of teaching were also similar between treatment and control teachers in the RCT sites, a result that differs from other research identifying participation in induction and having a mentor as related to higher retention (Borman & Dowling, 2008). On the whole, the retention rates for teachers in the RCT study (79 percent for treatment and 78 percent for control) were lower than those found among a national sample of teachers beginning teaching in 2007–08, among whom 85 percent remained in teaching 3 years later (Gray & Taie, 2015). This difference raises the possibility that local factors and/or more recent trends may be influencing retention patterns that induction might not address.

We did find significant and positive impacts on student achievement in ELA and mathematics after 2 years of NTC induction support for the combined RCT sites. These results suggest that the NTC induction model can improve the ELA and mathematics achievement of students in beginning teachers' classrooms. The QED using a differences-in-differences approach did not bear out positive impacts on student outcomes. The sample size of beginning teachers teaching ELA or mathematics in grades 4 through 8 that resulted from the participating districts' hiring patterns and testing schedules was very small, and the QED was extremely constrained in being able to detect any effects. The positive impacts found in the RCTs presented here contrast with a 2010 study of comprehensive teacher induction that reported no effects on student achievement after 2 years of induction support (Glazerman et al., 2010). In that study, effects in the third year of teaching, lagging the induction period, were positive and statistically significant but inconsistent under different approaches to estimating impact. Although we tested different scenarios than in the 2010 study, in all cases the student achievement impacts remained similar across the sensitivity analyses we conducted, indicating these findings are robust and stable (see Appendix G).

The results from this evaluation of the i3 Validation grant point to several implications for NTC's i3 Scale Up grant and for the field more broadly.

- High risk of attrition among the observation sample (due to both burden and turnover) means that the initial sample needs to account for an attrition rate higher than average turnover in the profession.
- While classroom observation tools such as the Framework for Teaching and the Classroom Assessment Scoring System (CLASS) have achieved reliability as measures for use in research studies, they may still apply to broad aspects of classroom practice that might not reflect instructional dimensions impacted by the NTC model on beginning teachers. For

instance, measures of how beginning teachers interrogate their own practice, how they make instructional decisions using data, how they think through integrating strategies for diverse learners, or how they exemplify mindsets of a reflective practitioner might round out the aspects of teacher practice NTC aims to improve.

- Developing the logic model and defining the implementation fidelity indicators provided an opportunity for NTC to clarify the aspects of the model program developers believed would be most meaningful in achieving results, and the indicators provided NTC a tool with which to establish common expectations and monitor progress with local program leaders.
- The NTC model by design is a comprehensive one supporting multiple levels of the system—teacher, mentor, school leader, and district. As NTC continues engaging districts across the country in its induction strategy, local contexts that vary at each of those levels will pose new opportunities to refine and adapt the model and raise new questions about which components are nonnegotiable and which can tolerate more flexibility and to what effect.

Under an i3 Scale Up grant that began in 2016, NTC is currently implementing its model in five urban districts across the country. SRI is conducting RCTs in each district. While NTC successfully achieved high implementation fidelity under the i3 Validation grant, scaling up to more districts and more diverse contexts necessitated adaptations to enhance sustainability and applicability. For example, full-time release mentors can be cost-prohibitive for some districts; the Scale Up sample includes several sites using school-based mentors who are retaining their own classroom full or part time as a more cost-efficient staffing model. Additionally, classroom observations as one of the primary sources of data by which mentors support their beginning teachers in examining instructional practice has practical limits if the observations need to be conducted in person. NTC integrated a classroom video tool into Learning Zone to promote more efficient and more frequent observation and feedback by mentors. The expectations for beginning teachers' induction experiences—frequency, intensity, content—have not changed, however. The i3 Scale Up study will build on the results reported here to determine whether and to what extent the NTC induction model with some specific adaptations can achieve high implementation fidelity in larger and more diverse district settings and whether, across these varying contexts, it has positive effects on teacher practice, teacher retention, and student achievement.

REFERENCES

Borman, G., & Dowling, N. M. (2008, September). Teacher attrition and retention: A meta-analytic and narrative review of the research. *Review of Educational Research, 78*(3), 367–409.

Danielson, C. (2013). *The framework for teaching evaluation instrument: 2013 edition*. Princeton, NJ: The Danielson Group.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study* (NCEE 2010-4028). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gray, L., & Taie, S. (2015). *Public school teacher attrition and mobility in the first five years: Results from the first through fifth waves of the 2007–08 Beginning Teacher Longitudinal Study* (NCES 2015-337). U.S. Department of Education. Washington, DC: National Center for Education Statistics. http://nces.ed.gov/pubsearch

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*, 84–117.

Hobson, A. J., Ashby, P., Malderez, A., & Tomlinson, P. D. (2009). Mentoring beginning teachers: What we know and what we don't. *Teaching and Teacher Education, 25*(1), 207–216.

Ingersoll, R., & Strong, M. (2011, June). The impact of induction and mentoring programs for beginning teachers: A critical review of the research. *Review of Education Research, 81*(2), 201–233.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M.D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms*. (NCSER 2013–3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences (IES), U.S. Department of Education. This report is available on the IES website, http://ies.ed.gov/ncser/

National Commission on Teaching and America's Future. (1997). *Doing what matters most: Investing in quality teaching.* New York, NY: Author.

National Commission on Teaching and America's Future. (2016). *What matters now: A new compact for teaching and learning.* New York, NY: Author.

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

Smith, T., & Ingersoll, R. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American Education Research Journal, 41*(3), 681–714.

Snyder, J., & Bristol, T. J. (2015). Professional accountability for improving life, college, and career readiness. *Education Policy Analysis Archives*, 23(16). Retrieved from http://files.eric.ed.gov/fulltext/EJ1070474.pdf

Wang, J., & Odell, S. (2002, Autumn). Mentored learning to teach according to standards-based reform: A critical review. *Review of Educational Research, 72*(3), 481–546.

APPENDICES
(*AVAILABLE IN A SEPARATE FILE WITH THE FULL REPORT AND ALL APPENDICES*)

SRI Education

SRI Education, a division of SRI International, is tackling the most complex issues in education to identify trends, understand outcomes, and guide policy and practice. We work with federal and state agencies, school districts, foundations, nonprofit organizations, and businesses to provide research-based solutions to challenges posed by rapid social, technological and economic change. SRI International is a nonprofit research institute whose innovations have created new industries, extraordinary marketplace value, and lasting benefits to society.

Silicon Valley
(SRI International headquarters)
333 Ravenswood Avenue
Menlo Park, CA 94025
+1.650.859.2000

Washington, DC Metro Area
1100 Wilson Boulevard, Suite 2800
Arlington, VA 22209
+1.703.524.2053

education@sri.com

www.sri.com/education