AIR®
AMERICAN INSTITUTES FOR RESEARCH®

Effects of National Board Certified Instructional Leaders on Classroom Practice and Student Achievement of Novice Teachers

A Study Report Developed for the National Board for Professional Teaching Standards

MAY 2019

Bo Zhu | Natalya Gnedko-Berry | Trisha Borman | David Manzeske

MAKING RESEARCH **RELEVANT**

AMERICAN INSTITUTES FOR RESEARCH | AIR.ORG

Effects of National Board Certified Instructional Leaders on Classroom Practice and Student Achievement of Novice Teachers

A Study Report Developed for the National Board for Professional Teaching Standards

MAY 2019

Bo Zhu | Natalya Gnedko-Berry | Trisha Borman | David Manzeske

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

www.air.org

8257_05/19

Contents

Page

Abstract 1

Introduction 2

Current Study 4

Methods 5

Study Design 5

Study Samples 5

Data Sources & Procedure 10

Outcome Measures 11

Analytic Approach 12

Results 13

Differences in Classroom Practice Between Intervention and Comparison Teachers Were Not Statistically Significant 13

Students in the Intervention Group Had Significantly Higher Achievement Than Students in the Comparison Group 15

Discussion 17

Limitations and Directions for Future Research 20

References 22

Appendix A. Characteristics of Original and Analytic Samples for the Analysis of Teacher Classroom Practice A1

Appendix B. Standardized Mean Differences Between Intervention and Comparison Groups at Baseline B1

Appendix C. CLASS Data C1

Appendix D. Statistical Approach to Estimate Differences in Teacher Classroom Practice D1

Appendix E. Statistical Approach to Estimate Differences in Student Achievement E1

Appendix F. Regression Results F1

Appendix G. Results of Sensitivity Analysis for Student Achievement G1

Tables

Page

Table 1. Steps for the Creation of the Teacher Analytic Sample 6

Table 2. Case Removal Process for Student Sample 8

Table 3. Numbers of Schools, Teachers, and Classes in ELA and Mathematics Samples 9

Table 4. Estimated Differences in Teacher Classroom Practice Between Intervention and Comparison Teachers 15

Table 5. Estimated Difference in General Achievement Between Intervention and Comparison Students 16

Table A1. Average CLASS Score and Characteristics of the Original Sample and Analytic Sample A2

Table B1. Standardized Mean Differences Between Intervention and Comparison Teachers at Baseline B1

Table B2. Standardized Mean Differences Between Intervention and Comparison Students at Baseline B2

Table C1. Description of CLASS Domains and Dimensions Shared Across All Three Levels C1

Table C2. CLASS Score Descriptives: Grades K–3, Raw Sample Mean & Standard Deviation From Manual C2

Table C3. CLASS Score Descriptives: Upper Elementary (Grades 4–6), Raw Sample Mean & Standard Deviation From Manual C3

Table C4. CLASS Score Descriptives: Secondary (Grades 7–12), Raw Sample Mean & Standard Deviation From Manual C3

Table F1. Estimated Differences in Teacher Classroom Practice Between Intervention and Comparison Teachers F1

Table F2. Estimated Differences in Student General Achievement Between Intervention and Matched Comparison Students Using Different Covariates F2

Table G1. Estimated Differences in Student General Achievement Between Intervention and Matched Comparison Students Using Different Covariates G1

Table G2. Estimated Differences in Student General Achievement Between Intervention and Matched Comparison Students Using Different Covariates G1

Figures

Page

Figure 1. Standardized Average CLASS Scores of Intervention and Comparison Teachers By Domain 14

Figure 2. Standardized Average Test Scores of Intervention and Comparison Students 16

Abstract

The study examined the effect of National Board Certified Teachers (NBCTs) in instructional leadership roles, operationalized as mentors to novice teachers, on (a) classroom practices of mentored novice teachers in Grades K–12 and (b) student achievement of mentored teachers' students in Grades 4–8. The study compared outcomes between NBCT mentors and non-NBCT mentors. The study examined the effect of NBCT mentors after one academic year and was conducted in San Francisco Unified School District. Using the Classroom Assessment Scoring System, we examined novice teachers' classroom practices on the domains of Emotional Support, Classroom Organization, Instructional Support, and across all three domains. The results did not reach statistical significance, but the effect sizes for Emotional Support, Classroom Organization, and a global measure across all domains suggest meaningful differences between classroom practices of novice teachers mentored by NBCTs and non-NBCTs. These effect sizes were 0.28, 0.28, and 0.21 standard deviations, respectively. The effect size for the domain of Instructional Support was near zero at –0.06 standard deviations. Our sample size for the analysis of classroom practices did not have sufficient power to estimate differences at a statistically significant level. We examined student achievement using the state's standardized test scores in mathematics and English language arts. Our achievement measure includes either subject: That is, we did not estimate effects separately for mathematics and English language arts. The results suggest that students taught by teachers mentored by NBCTs had a higher level of achievement than students mentored by non-NBCTs. The difference was statistically significant at a *p* value of .05, and the effect size was meaningful at 0.18 standard deviations. Small sample sizes and low statistical power prevent us from making confident conclusions about the effect of NBCTs in instructional leadership roles on classroom practices of supported teachers and student achievement. However, the evidence is encouraging and warrants additional rigorous research on the impact of NBCTs as instructional leaders.

Introduction

The interest in teacher leadership has grown steadily in the past two decades. The term "teacher leadership" encompasses many instructional roles, including mentors, coaches, curriculum support providers, teaching specialists, or teachers on special assignment (Silva, Gimbert, & Nolan, 2000; York-Barr & Duke, 2004). Teachers who fill these roles are expected to use classroom expertise and leadership skills to facilitate professional growth of their colleagues, ultimately contributing to improvement in student learning (e.g., Backes & Hansen, 2018; Campbell & Malkus, 2011; Ingersoll & Strong, 2011; Jackson & Bruegmann, 2009; Kraft, Blazar, & Hogan, 2018; Mihaly, Master, & Yoon, 2015; York-Barr & Duke, 2004; Yuan, 2015).

With increased interest in teacher leadership, the body of research in this area has grown. Studies suggest that teachers learn well from other teachers (e.g., Evertson & Smithey, 2000; Stanulis & Floden, 2009; Thompson, Paek, Goe, & Ponte, 2004), even when teacher leaders are not serving in formal roles (e.g., Yuan, 2015). Studies also suggest that teacher leadership could have a positive effect on teacher leaders' own instructional practice and professional growth (e.g., Ryan, 1999; York-Barr & Duke, 2004). Studies that examined the effect of teacher leadership on student outcomes produced inconclusive findings; some studies suggest a positive relationship between teacher leadership and student outcomes (e.g., Campbell & Malkus, 2011) and others suggest no relationship or mixed results (e.g., Mihaly, Master, & Yoon, 2015).

A larger body of research on teacher leadership has examined teacher induction—a form of teacher leadership in which experienced teachers mentor novice teachers. Research suggests a positive effect of teacher induction on novice teachers' retention rates (e.g., Cohen & Fuller, 2006; Fuller, 2003; Ingersoll & Smith, 2004; Kapadia, Coca, & Easton, 2007; Strong, 2005; Strong, 2009; Tushnet et al., 2002) and use of effective classroom practices (e.g., Evertson & Smithey, 2000; Thompson et al., 2004). Some studies also suggest that students of novice teachers who received mentorship have better academic outcomes than students of novice teachers who were not mentored (e.g., Glazerman et al., 2010; Strong, 2006). As with other areas of teacher leadership, research on the impact of teacher induction on teacher and student outcomes has reported mixed results. For example, a randomized controlled trial of a teacher induction program did not find evidence of a positive effect of induction on mentored teachers' retention rates or their classroom practice, but the results suggested a positive impact on the achievement of students who were taught by mentored teachers (Glazerman et al., 2010).

Growing interest in teacher leadership has led to an increase in the investment in teacher leaders, with school districts and other educational agencies and nonprofit organizations allocating increasing resources to the development and support of teacher leaders (e.g., Center for Strengthening the Teaching Profession, 2016; York-Barr & Duke, 2004). With many pathways to teacher leadership becoming available (e.g., through training provided by schools and districts or leadership courses through external organizations [Association for Supervision and Curriculum Development, 2014; Levenson, 2014]), it may be difficult for teachers to decide which path is right for them. It also may be difficult for schools and districts to decide which path they should support to identify and deploy teacher leaders effectively.

Teacher certification offered by the National Board for Professional Teaching Standards (National Board) may be a useful solution for districts and schools interested in teacher leadership for two reasons: (1) National Board Certified Teachers (NBCTs) combine characteristics expected of teacher leaders and (2) some school districts already have access to NBCTs. The characteristics of NBCTs that are expected of teacher leaders include effectiveness in the classroom (Cantrell, Fullerton, Kane, & Staiger, 2008; Cavalluzzo, 2004; Cowan & Goldhaber, 2016; Goldhaber & Anthony, 2007; Harris & Sass, 2009; Manzeske et al., 2017) and the ability to provide instructional leadership to other teachers (Belson & Husted, 2015; Cannata, McCory, Sykes, Anagnostopoulos, & Frank, 2010; Harris & Sass, 2009; Loeb, Elfers, Plecki, Ford, & Knapp, 2006; National Board, 2016; Sato, Hyler, & Monte-Sano, 2014). For example, Cowan and Goldhaber (2016) found that NBCTs were more effective than non-NBCTs in reading and mathematics in Grades 4–8. Harris and Sass (2007, 2009) found that when the number of NBCTs at a school serving in formal mentorship roles increased, so did student achievement in mathematics and reading. As far as access to NBCTs, approximately 122,000 teachers in the United States are already certified, and thousands more are working on certification (National Board, 2018). Existing and future NBCTs provide a pool of teachers that schools and districts investing in National Board certification could readily use to identify prospective teacher leaders.

Although NBCTs have the qualities associated with teacher leadership and are employed in school districts in instructional leadership roles, no causal studies have examined the effectiveness of NBCTs as instructional leaders. For example, the study by Harris and Sass (2007, 2009) was correlational and did not directly test the relationship between NBCT mentors and student achievement of teachers they mentored. Other studies also used descriptive or correlational designs without testing the direct link between NBCT instructional leaders and teacher and student outcomes (e.g., Belson & Husted, 2015). The current study begins to address this gap in knowledge by using a quasi-experimental design (Shadish, Cook, &

Campbell, 2002) to study the effect of NBCT instructional leaders on classroom practice of mentored teachers and student achievement of mentored teachers.

Current Study

The current study examined the effect of NBCT instructional leaders in mentorship roles on teacher and student outcomes. The study addressed the following two research questions:

1. What is the effect of NBCT mentors on the classroom practice of novice teachers in Grades K–12?
2. What is the effect of NBCT mentors on the achievement of mentored teachers' students in Grades 4–8?

The grade levels addressed by each question were determined by data availability, which we describe later in the report. We examined the effectiveness of NBCT mentors after one school year—the first year of novice teachers' independent teaching. We conducted the study in 2015–16 and 2016–17 in San Francisco Unified School District (SFUSD). The district offers an induction program, Beginning Teacher Support and Assessment (BTSA), to all novice teachers. BTSA is a state-funded induction program that supports novice teachers in completing requirements for the California Clear Multiple Subjects, Single Subjects, and Educational Specialist credentials. The program provides all novice teachers 2 years of support, including coaching, professional development, and formative assessment using observations and teaching portfolios, toward credential completion (SFUSD, 2018).

Approximately 300 new teachers in San Francisco participate in BTSA each year (half are in their first year of the program, and half are in their second year), and approximately 100 mentors support them. BTSA program planners carefully select mentors through a formal application process. For example, mentors are supposed to have a minimum of 3 years of effective teaching experience and have knowledge of the content area of the novice teacher's teaching assignment. They are expected to provide no less than 1 hour per week of individualized support to novice teachers. BTSA mentors are matched with new teachers to ensure the mentor has the needed subject expertise and to allow for geographic proximity of mentors and mentees. Usually, mentors support one or two novice teachers. Selected mentors receive training from BTSA on the effective ways of coaching and mentoring, goal setting for the novice teachers they support, best practices in adult learning, individual support and reflection on mentoring practices, and ways to navigate BTSA standards and requirements (SFUSD, 2015).[1]

[1] BTSA information was retrieved from the SFUSD website and gathered from conversations with the program providers during the study's implementation.

BTSA mentors include NBCTs and non-NBCTs. The involvement of NBCT and non-NBCT mentors in BTSA provided the needed conditions for the current study: We could observe and compare the outcomes of novice teachers mentored by NBCTs and non-NBCTs to answer the study's research questions. BTSA does not require mentors to be NBCTs nor does it expect NBCT mentors to provide support to novice teachers that is different from the support provided by non-NBCT mentors. However, given the explicit focus of the National Board Certification process on teacher leadership skills, we hypothesized that the impact of NBCTs on novice teachers would be different compared with non-NBCTs. This hypothesis served as an impetus for the current study.

In the following sections, we describe the study's methods and results. We conclude with a discussion of findings.

Methods

In the following section we describe the details of study design, samples, data sources, and analytic approach.

Study Design

We used a quasi-experimental design to examine the effect of NBCT mentors on classroom practice of novice teachers (Research Question [RQ] 1) and the effect of NBCT mentors on student achievement of supported teacher (RQ 2). For RQ 1, the intervention group consisted of novice teachers supported by NBCT mentors, and the comparison group consisted of novice teachers supported by non-NBCT mentors. All teachers in both groups were participating in BTSA. For RQ 2, the intervention group consisted of students taught by teachers mentored by NBCTs, and the comparison group consisted of students taught by teachers mentored by non-NBCTs. We examined outcomes after one academic year of mentorship, which was the first year of teachers' independent teaching and participation in BTSA.

Study Samples

We used two samples: a teacher sample to examine classroom practice (RQ 1) and a student sample to examine student achievement (RQ 2). Below we describe how we created each sample.

Teacher Sample for the Analysis of Classroom Practice

The initial teacher sample included 111 teachers from 52 schools: 28 in the intervention group (mentored by NBCTs) and 83 in the comparison group (mentored by non-NBCTs). These were

all teachers in their first year of receiving mentorship who were observed at least once in 2015–16 or 2016–17. We combined 2015–16 and 2016–17 cohorts to maximize statistical power (we confirmed with program planners that the implementation of BTSA was the same in both years). We removed four teachers whose data were missing from the districts' administrative records (for example, missing information on demographic characteristics, school, and grade assignment). Next, we removed 47 teachers for whom baseline or outcome data on classroom practice were missing (measured by the Classroom Assessment Scoring System [CLASS] score described in Data Sources). The final sample of teachers, therefore, included 60 teachers with both baseline and outcome data on classroom practice (i.e., complete cases): 12 in the intervention group and 48 in the comparison group. Teachers included in the sample taught grades K-12. Table 1 summarizes all steps for the creation of the teacher analytic sample.

Table 1. Steps for the Creation of the Teacher Analytic Sample

Steps	Intervention Teachers	Comparison Teachers	Total
Initial Sample of Teachers	**28**	**83**	**111**
Step 1: Removed teachers: missing district administrative data	0	4	4
Step 2: Removed teachers: missing either baseline or outcome CLASS score	16	31	47
Analytic Sample of Teachers	**12**	**48**	**60**

We examined differences between teachers in the final analytic sample and teachers who were removed from the sample on three sets of characteristics: classroom practice, demographics, and school characteristics. All differences between teachers in the analytic sample and teachers who were removed from the sample were not statistically significant, suggesting that data missingness was random. Detailed results are presented in Appendix A.

Next, we examined baseline equivalence on the outcome measures (i.e., teachers' classroom practice) between the intervention and comparison teachers included in the analytic sample. Because the outcome measures were continuous, we used Hedges' *g* to compute standardized mean differences with the small sample correction. The differences were between 0.05 and 0.25 standard deviations (SDs), suggesting that the groups were similar on baseline measures of classroom practice but that a statistical adjustment was needed in the analysis, which we implemented (What Works Clearinghouse, 2017).

We also examined baseline differences between the groups on additional teacher and school characteristics using Hedges' *g* to compute standardized mean differences for continuous variables and the Cox index for binary variables. Some differences were within the 0.05–0.25 SD range (e.g., race/ethnicity, full-time status), and some differences exceeded this range (e.g., gender, grades taught, school size). Because the additional characteristics were not central to our analysis (i.e., they were not baseline outcome measures), we decided against matching that would have improved balance on these characteristics but also would have reduced our sample size and, therefore, power to detect impact. Instead, we used statistical adjustments in the analysis for all covariates. See Appendix B for a summary of baseline characteristics and standardized mean differences between intervention and comparison teachers on these characteristics.

Student Sample for Analysis of Student Achievement

The initial student sample included 6,901 students taught by at least 1 of 100 teachers from the teacher sample (11 teachers from the teacher sample were not linked to any students in student achievement data provided by the district[2]). These students were in grades PK–12. We limited the sample to students in Grades 4–8 to ensure that they were in tested grades and had prior year test scores (test scores were our measure of student achievement, which we describe in Data Sources). For the same reason, we limited the sample to students who had scores in English language arts (ELA) or mathematics (science tests are not given in each grade, which is why prior year test scores were not available for this subject). We further limited the sample to those students whose teachers had baseline CLASS scores. We implemented this limitation because teachers' classroom practice can vary greatly when they begin teaching due to differences in the preparation program, student teaching, school and classroom assignments, etc. Therefore, we decided it was necessary to control for teachers' baseline classroom practice in the analysis.

A subset of students in our sample received instruction from both treatment and comparison teachers (i.e., novice teachers mentored by NBCTs and non-NBCTs). For these students, we removed records linked to teachers mentored by non-NBCTs and kept a randomly selected achievement record linked to teachers mentored by NBCTs. Our incentive for keeping records linked to teachers mentored by NBCTs was to maintain as many students and teachers as possible in the intervention group, which was smaller relative to the comparison group. As a result, each student in our sample was linked to only one teacher within a subject.

[2] The district could not locate any course information for these teachers.

We implemented matching to construct a comparison group of students, which we describe later in this section. We mention it here because matching prompted further removal of students. Because we matched within subject and grade levels, we could only keep students in our sample if we had a pool of students in the same subject and grade to whom we could match them. We removed students who did not have a pool of matches. After all these steps were taken to refine the sample, we were left with 303 unique students in Grades 4–8 with baseline and outcome achievement data: 121 students in the intervention group (taught by 3 teachers mentored by NBCTs) and 182 students in the comparison group (taught by 4 teachers mentored by non-NBCTs). Table 2 summarizes all steps we took to refine the student sample.

Table 2. Case Removal Process for Student Sample

	Number of Students
Original data set with students linked to at least one teacher from the teacher sample	**6,901**
Step 1: Removed students in grades other than 4–8	5,120
Step 2: Removed students in classes other than ELA/mathematics	856
Step 3: Removed students missing baseline or outcome achievement data	59
Step 4: Removed students whose teachers were missing baseline CLASS scores	141
Step 5: For students in both intervention and comparison classrooms, removed records from comparison classrooms and randomly chose one record from an intervention classroom	31
Step 6: Removed students from subjects and grades taught by only treatment or comparison teachers	391
Remaining student sample	**303**

Of the 303 unique students we identified in the previous steps, 58 fourth graders had ELA achievement data and were in our initial ELA sample. The remaining 245 sixth and seventh graders had mathematics achievement data and were in our initial mathematics sample.

In the sample of 58 students with ELA achievement data, 28 were in the intervention group (taught by teachers mentored by NBCTs); the remaining 30 students were the pool for creating a comparison group. In the sample of 245 students with mathematics achievement, 93 were in the intervention group (taught by teachers mentored by NBCTs); the remaining 152 students were the pool for creating a comparison group. We decided to use matching to create a comparison group of students because our initial review of data suggested that students taught

by NBCT and non-NBCT mentors were different at baseline on achievement. We used propensity score matching to identify comparison students who were most similar to intervention students based on prior student achievement and additional background characteristics: English learner (EL) status, special education (SPED) status, gender, and racial/ethnic minority status.[3] The matching process was conducted separately for each combination of grade and subject. Each student who received the intervention was matched with one student who did not receive the intervention based on the closest propensity score (i.e., the student with the closest propensity to being in an intervention teacher's classroom). We conducted matching with replacement, meaning that a matched student was added back to the pool of possible matches. Therefore, some comparison students were matched with multiple intervention students. We used weights to control for the number of times a student was matched. Once matched, we combined students in Grades 6 and 7 for the mathematics sample. We did not need to combine students for the ELA sample because they were all from Grade 4. Table 3 shows numbers of students included in each condition (i.e., intervention or comparison) by subject, along with numbers of classes, teachers, and schools.

Table 3. Numbers of Schools, Teachers, and Classes in ELA and Mathematics Samples

Subgroup	Number of Students	Number of Classes	Number of Teachers	Number of Schools
ELA—Intervention	23	2	2	2
ELA—Comparison	9	1	1	1
Mathematics—Intervention	93	2	1	1
Mathematics—Comparison	48	4	3	3

Note. ELA = English language arts

Next, we combined ELA and mathematics samples. We made this decision to remedy the $n = 1$ confounding factor at the teacher level that existed in individual samples (What Works Clearinghouse, 2017). That is, the ELA sample included one teacher in the comparison group, and the mathematics sample included one teacher in the intervention group (see Table 3). Therefore, any estimates of the effectiveness would have been confounded by only one teacher per condition. By combining the ELA and mathematics samples, we created a final analytic sample that included 116 intervention students and 57 comparison students. Because we

[3] The district did not share FRPL information with external researchers.

combined the ELA and mathematics samples, we used the pooled *z*-scores in ELA and mathematics as a general measure of achievement.

We examined baseline equivalence for the final student sample. Using Hedges' *g*, we examined group differences on the measure of general achievement between intervention and comparison students. The difference was 0.04 SDs, suggesting that the groups were similar on achievement at baseline (What Works Clearinghouse, 2017). We also examined baseline equivalence for other student characteristics. We used Hedges' *g* to compute standardized mean differences for continuous variables and the Cox index for binary variables. Some differences were within the 0.05–0.25 SD range (e.g., student gender, SPED status in both samples), and some differences exceeded this range (e.g., EL status, student race/ethnicity). Because the additional student characteristics were not central to our analysis (i.e., they were not baseline outcome measures), we decided against further matching that could have improved balance on these characteristics but also would have reduced our sample size and, therefore, power to detect impact. Instead, we used statistical adjustments in the analysis for all student covariates. See Appendix B for a summary of baseline characteristics and standardized mean differences between intervention and comparison students on these characteristics.

Data Sources & Procedure

Our data sources included classroom observations and administrative records for teachers, students, and schools.

Classroom Observations

We used CLASS scores to measure teachers' classroom practices. CLASS has been validated through field testing across thousands of classrooms from preschool to secondary grades (Kane & Staiger, 2012; Pianta, Hamre, & Mintz, 2012a, b). CLASS assesses quality of student–teacher interactions on four domains: Emotional Support, Classroom Organization, Instructional Support, and Student Engagement (Pianta, La Paro, & Hamre, 2008; Pianta, Hamre, & Mintz, 2012a, b). Each domain comprises one to five dimensions, each of which specifically measures one aspect of student–teacher interactions. The study only included the first three domains[4] shared across the CLASS grade levels (K–3, upper elementary, and secondary). A global measure is generated based on the three domains. Appendix C includes additional information on CLASS domains.

[4] CLASS K–3 does not assess the quality of student–teacher interaction in the student engagement domain. Because our analysis spans all grades, we did not include this domain.

American Institute for Research (AIR) researchers trained and certified as CLASS observers conducted in-person observations. For the first cohort of teachers, observations were conducted in fall 2015 (baseline) and spring 2016 (outcome). For the second cohort, observations were conducted in fall 2016 (baseline) and spring 2017 (outcome). AIR researchers recruited intervention and comparison teachers for participation in classroom observations through direct e-mail contact. BTSA program planners provided teachers' contact information to AIR.

Administrative Records

Administrative records collected by SFUSD for grades K-12 were a source of data on student achievement, student and teacher characteristics, and school characteristics. Student data included students' test scores, demographics and grade and classroom information. Teacher data included school assignment, grade and subject assignments, and demographic characteristics. School data included school size and demographic characteristics of student body. Appendix B includes a summary of variables for the study's samples.

Outcome Measures

We used classroom observation scores as an outcome measure of teachers' classroom practice and standardized test scores (in mathematics and English Language Arts) as an outcome measure of student achievement.

Teacher Classroom Practice Was Measured Using Classroom Observations

We used four scores to measure classroom practice. Three scores reflected CLASS domains of teacher practice: Emotional Support, Classroom Organization, and Instructional Support. The fourth score was the global score, which combined results across domains. To compute domain scores, we first realigned dimensions across grade levels to ensure that dimensions within each domain were the same across all three grade levels[5]. We took this step because our analysis of outcomes combines across grade levels. The Emotional Support domain in our analysis includes Positive Climate, Negative Climate, Teacher Sensitivity, and Regard for Student/Adolescent Perspectives.[6] The Classroom Organization domain includes Behavior Management, Productivity, and Instructional Learning Formats. Finally, Instructional Support includes Quality

[5] Teachstone, who oversees the use of CLASS, provided permission to AIR to re-align the dimensions consistently across grade levels. Specifically, for upper elementary and secondary grade levels, we moved Negative Climate and Instructional Learning Formats dimensions from the Classroom Organization domain to the Emotional Support domain.

[6] Teachstone indicated that there was no difference between Regard for Student Perspectives (K–3 and Upper Elementary) and Regard for Adolescent Perspectives (Secondary).

of Feedback (see Appendix C for additional information on CLASS dimensions). Dimensions were scored on a 7-point scale, with 7 indicating the highest level of classroom practice. Dimensions have distinct score distributions by grade level. Therefore, we standardized scores against the means reported in the CLASS manuals.[7] [8] We computed domain scores by averaging across standardized dimension scores within each domain. We computed global score by averaging across domain scores. Further information on the standardization process can be found in Appendix C.

Student Achievement Was Measured Using Standardized Test Scores

Student academic achievement was measured using scale scores from the Smarter Balanced Assessment in either English language arts (ELA) or mathematics. We report the impact on general achievement, defined as an estimate pooled across ELA and mathematics. We did not estimate impact separately for ELA or mathematics because of subject-specific sample size limitations that we discussed in detail in Study Samples (subject-specific samples introduced an $n = 1$ confound). To facilitate pooling of samples across cohorts, grades, and subjects, we converted students' test scores into standardized scores (z-scores). Standardized scores were computed as the difference between a student's scaled score and the original sample's average scaled score for the assessment (in a given subject, grade, and year), divided by the sample's standard deviation of raw scores for that assessment.

Analytic Approach

Analyses of Teacher Classroom Practice

To address RQ 1, we used a two-level regression analysis that accounted for teachers nested in schools. We examined differences between classroom practice of teachers mentored by NBCTs and teachers mentored by non-NBCTs on the three key domains of Emotional Support, Classroom Organization, and Instructional Support. We also examined differences based on the

[7] Scores are standardized to the populations provided in each level's CLASS manual:
Pianta, R., LaParo, K., Hamre, B. (2008). *Classroom Assessment Scoring System manual: K–3.* Baltimore: Paul H. Brookes Publishing Co.
Pianta, R., Hamre, B., Mintz, S. (2012). *Classroom Assessment Scoring System manual: Upper elementary.* Charlottesville, VA: Teachstone.
Pianta, R., Hamre, B., Mintz, S. (2012). *Classroom Assessment Scoring System manual: Secondary.* Charlottesville, VA: Teachstone.

[8] Overall, our sample means are similar to the means reported in the CLASS manuals. Our means were slightly higher than the means in CLASS manuals for the following dimensions and grade levels: Positive Climate and Negative Climate dimensions in secondary grade level, Teacher Sensitivity dimension across all grade levels, Regard for Student/Adolescent Perspectives in upper elementary and secondary levels, Productivity and Instructional Learning Formats dimension in K–3 and secondary levels, and Quality of Feedback in K–3 level.

global score. We ran a separate model to examine each outcome. All analyses controlled for teacher background characteristics (i.e., gender, race/ethnicity, full-time employment status, and grade levels taught) and school background characteristics (i.e., total school enrollment, EL students, students with disability, percent female, and FRPL-eligible students). More technical details of the approach, such as the specification of the statistical model, are discussed in Appendix D.

Analyses of Student Achievement

We used a single-level regression analysis to examine the final sample. We chose this approach instead of a multilevel regression because the small number of clusters in our sample did not support the use of a multilevel regression[9] (see Table 3). The analysis controlled for relevant student characteristics (e.g., prior achievement), baseline measures of teacher classroom practice, and weights for comparison students to balance students who were matched more than once. [10] Comparison students with a greater propensity for being in an intervention (comparison students who were matched to more than one intervention student) were given more weight than those with a lower propensity (comparison students who were matched fewer times), controlling for the proportion of treatment and comparison students in the sample. Additional details about the analysis of student achievement, including the specification of the statistical model, are discussed in Appendix E.

Results

We report the results of analyses separated by research questions. First, we report the results of analysis that examined differences in classroom practice of intervention teachers (novice teachers mentored by NBCTs) and comparison teachers (novice teachers mentored by non-NBCTs). Second, we report the results of analysis that examined differences in achievement between intervention students (taught by novice teachers mentored by NBCTs) and comparison students (taught by novice teachers mentored by non-NBCTs).

Differences in Classroom Practice Between Intervention and Comparison Teachers Were Not Statistically Significant

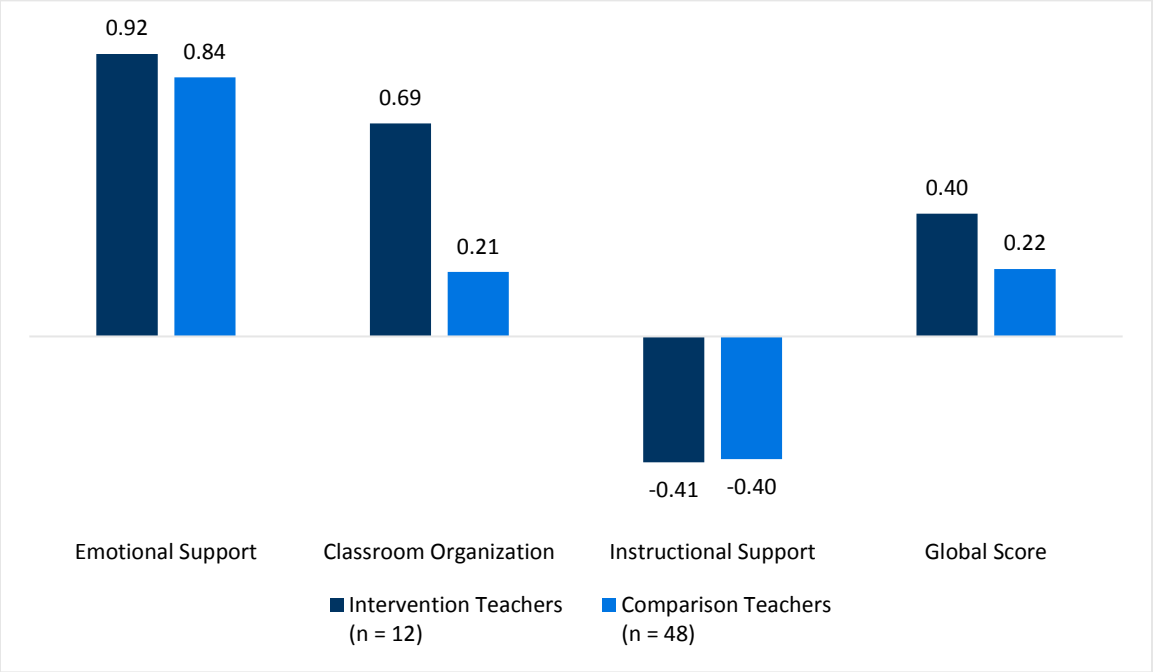
After 1 year of independent teaching, intervention teachers scored higher than comparison teachers on the Emotional Support and Classroom Organization domains of classroom practice: the average differences in standardized scores were 0.08 and 0.48, respectively. The

[9] For example, studies suggest that models with fewer than 20–25 clusters may not provide accurate estimates of the regression coefficients and standard errors (O'Dwyer & Parker, 2014).

[10] We did not include school-level covariates in the student analysis because (a) students came from a small number of schools with little variability between them, and (b) school-level covariates were correlated with student-level covariates.

intervention teachers scored lower than comparison teachers on the Instructional Support domain of classroom practice. The average difference in standardized scores was 0.01. The global score, which was the average across domains, was higher by 0.18 points for the intervention teachers. These results are shown in Figure 1. Note that these results are descriptive.

Figure 1. Standardized Average CLASS Scores of Intervention and Comparison Teachers By Domain



Note. The results are descriptive. The numbers are shown in standard deviations. Global scores were calculated by averaging scores between three domains (Emotional Support, Classroom Organization, and Instructional Support).

We used regression analysis to examine differences between the groups, controlling for baseline instructional practice and teacher and school characteristic. The results of this analysis are summarized in Table 4. Positive estimates indicate that intervention teachers scored higher than comparison teachers, whereas negative estimates indicate that intervention teachers scored lower. None of the estimates were statistically significant. However, the effect sizes suggest that differences between the intervention and comparison teachers could be meaningful. On the domains of Emotional Support and Classroom Organization the intervention teachers' scores were 0.28 SDs greater than the scores of comparison teachers. The global score for the intervention teachers was 0.21 SDs greater. The difference on Instructional

Support favored the comparison teachers, but that difference was very small at 0.06 SDs. More detailed results are presented in Appendix F.

Table 4. Estimated Differences in Teacher Classroom Practice Between Intervention and Comparison Teachers

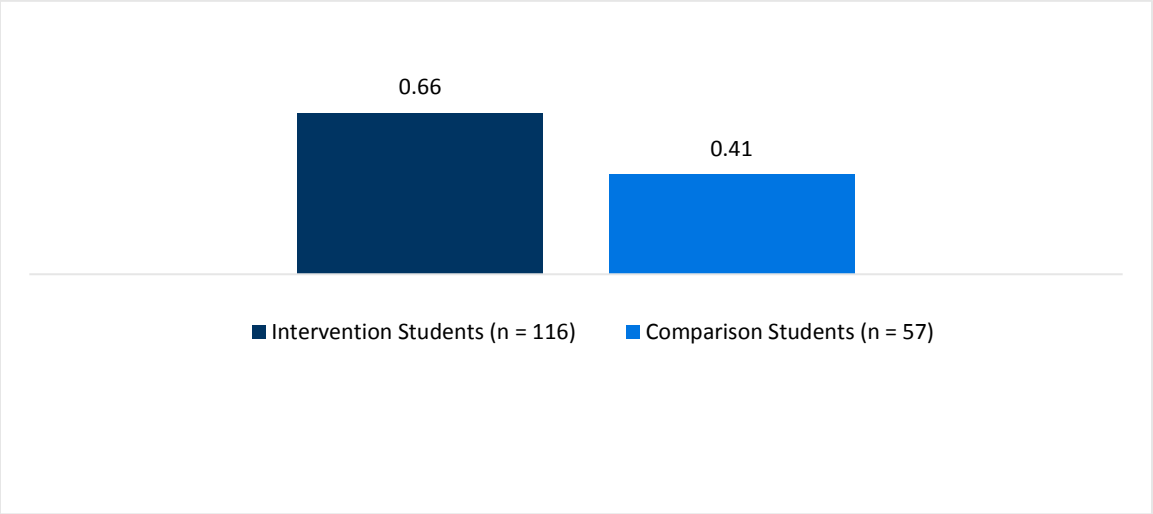
Domain	Estimates (Standard Error)	*p* Value	Effect Size
Emotional Support	0.367 (0.407)	.368	0.280
Classroom Organization	0.412 (0.426)	.333	0.281
Instructional Support	-0.102 (0.479)	.832	-0.061
Global score	0.266 (0.379)	.483	0.214

Note. The analyses included 12 intervention teachers and 48 comparison teachers. Standard errors are presented in parentheses. Effect sizes were computed using standardized mean differences (Hedges' *g*) with the small sample correction.

Students in the Intervention Group Had Significantly Higher Achievement Than Students in the Comparison Group

After 1 academic year, intervention students (taught by novice teachers mentored by NBCTs) had higher test scores than the comparison students (taught by novice teachers mentored by non-NBCTs). The average difference between the groups was 0.25 in standardized scores (see Figure 2 for descriptive findings).

Figure 2. Standardized Average Test Scores of Intervention and Comparison Students



Note. The results are descriptive. The numbers are shown in standard deviations.

We examined group differences using a regression model that controlled for teachers' baseline classroom practices using a global CLASS score and student characteristics. The results suggest that students who received the intervention (taught by novice teachers mentored by NBCTs) had a higher general achievement than the comparison students (taught by novice teachers mentored by non-NBCTs). The difference between the intervention and comparison groups of students was statistically significant at *p* value of .054. The effect size of difference between the groups was 0.18, suggesting that difference between the groups could be meaningful beyond the level of statistical significance. Table 5 summarizes the results of this analysis. More detailed results are presented in Appendix F.

Table 5. Estimated Difference in General Achievement Between Intervention and Comparison Students

	Estimates	***p* Value**	**Effect Size**
Intervention (controlling for teachers' global baseline CLASS score)	0.134* (0.069)	.054	0.178*

Note. Analysis included 116 intervention students and 57 comparison students and controlled for students' prior achievement and other student characteristics. The estimate is shown in standard deviations. Standard error is presented in parentheses. * $p \leq .05$.

We ran additional regressions that controlled for teachers' baseline classroom practices using CLASS scores in individual domains (Emotional Support, Classroom Organization, and

Instructional Support) instead of one global score. We ran these regressions as a sensitivity analysis. Except for the Emotional Support domain, the results were significant and in the same direction as the result from the model that used the global score (see Table G1 in Appendix G). When we controlled for Emotional Support, the differences between the groups were not statistically significant and the effect size was near zero at -0.03. Further exploration of the data suggests that the lack of difference in student achievement when we controlled for Emotional Support could be due to the distribution of scores among teachers on this domain. Teachers' scores in this domain had a substantial fluctuation, likely cancelling or diluting any differences between teachers and influencing the achievement estimate for that one model.

Discussion

NBCTs have the teaching and leadership skills expected of teacher leaders (e.g., Belson & Husted, 2015; Cowan & Goldhaber, 2016; Harris & Sass, 2009) and are employed in school districts nationwide in instructional leadership roles. To date, however, research has not directly examined the effectiveness of NBCTs as instructional leaders. The current study examined the effect of NBCT instructional leaders, operationalized as mentors to novice teachers, on the classroom practice of mentored teachers. The study also examined whether students of teachers who were mentored by NBCTs had a different level of achievement than students of teachers who were mentored by non-NBCTs. The length of mentorship was one academic year.

Our analysis of classroom practice shows that at the end of one academic year of mentorship, the differences between novice teachers mentored by NBCTs and novice teachers mentored by non-NBCTs were not statistically significant. The effect sizes for Emotional Support, Classroom Organization, and the global score across the three domains included in the analysis (Emotional Support, Classroom Organization, and Instructional Support) were 0.28, 0.28, and 0.21 SDs, respectively. These effect sizes are below the reported average. For example, a meta-analysis of teacher coaching found that the average effect size of coaching on classroom practice of supported teachers was 0.49 SDs (Kraft et al., 2018). This effect size, however, is based on interventions that compared teachers who received coaching with teachers who did not. Our study did not have a condition in which teachers did not receive mentorship: All teachers were mentored, but the intervention group of teachers was mentored by NBCTs. Therefore, the effect sizes in our study reflect the difference between two mentorship conditions (mentored by NBCT and non-NBCT), which may be why they are smaller than the average reported by Kraft et al. (2018). Furthermore, the study by Kraft et al. (2018) did not include teacher induction

programs that focused on novice teachers like the current study did, which could be another source of difference in the effect sizes. Our interpretation of the effect sizes found in the current study for Emotional Support, Classroom Organization, and the global score is that they suggest a meaningful contribution of NBCT mentors to the professional growth of novice teachers—a contribution that is above of what other experienced mentors in our study provided to their mentees.

The effect size for the difference between novice teachers mentored by NBCT and non-NBCT on the domain of Instructional Support was much smaller than the domains of Emotional Support and Classroom Organization, and in the opposite direction: –0.06 SDs. NBCT and non-NBCT mentors could have provided a similar support to their mentees in this domain, which would help explain why the effect size is near zero. The similarity of support on this domain between NBCT and non-NBCT mentors is likely because BTSA pairs novice teachers with mentors based on shared instructional area (i.e., mentors experienced in teaching science are likely to be paired with novice teachers assigned to teach science). Because of this pairing it is possible that all BTSA mentors, NBCT and non-NBCTS, had a similar approach to mentoring novice teachers in the areas that were scored in the Instructional Support domain, such as promoting concept development among their students. BTSA does not incorporate Emotional Support and Classroom Organization to mentor–mentee pairing, which lends further support to why the current study found larger differences in effect sizes on these domains but not on the domain of Instructional Support.

Our analysis of student achievement used a general measure of achievement as outcome (i.e., either an ELA or mathematics test score). We found that students taught by teachers who were mentored by NBCTs had a higher level of achievement than students whose teachers were mentored by non-NBCTs. The difference was statistically significant at *p* value of .05. The effect size difference between the groups was 0.18 SDs. This effect size is in line with the average effect size of teacher coaching on general student achievement reported by Kraft et al. (2018).[11] The difference of 0.18 SDs translates to 6.5 months of additional learning for students whose teachers were mentored by NBCTs,[12] assuming 9 months of instruction in a year. This increase in learning is substantial and especially notable because students in our study were taught by novice teachers, who tend to be less effective than their peers who have more experience (e.g., Henry, Fortner, & Bastian, 2012; Rockoff, 2004). It is important, however, to remember that students in the intervention group were taught by only three NBCT-mentored

[11] General student achievement in the Kraft et al. (2018) study is based on scores in mathematics, reading, and science.
[12] This translation was based on a typical annual growth of 0.25 SDs per year for all grades and subjects, estimated by Hanushek, Woessmann, and Peterson (2012).

teachers. Therefore, although this finding is promising, it lacks generalizability and needs to be examined through additional research, which we discuss in Limitations and Directions for Future Research.

The only published work we located that specifically focused on the relationship between NBCTs in instructional leadership roles and student achievement is the study by Harris and Sass (2007, 2009). The authors found that at schools with a greater number of NBCTs in official mentorship roles, the achievement of students taught by non-NBCTs significantly increased in both mathematics and reading but only on one test included in the study. The authors explained this finding as a positive spillover effect of NBCT mentors on the instructional practices of their noncertified colleagues. Results of the current study align with the results of the earlier study in that we also found a positive relationship between NBCT mentorship and student achievement. The effect size on student achievement in the current study is greater: 0.18 compared with 0.01 in mathematics and reading in the earlier study. However, the earlier study did not link NBCT mentors to supported teachers or supported teachers' students—the differences that likely contributed to the gap in the student achievement effect sizes seen in the current study and the study by Harris and Sass (2007, 2009). Because meta-analysis by Kraft et al. (2018) included studies that linked coaches to supported teachers and supported teachers' students, even though it didn't differentiate between the NBCT status of coaches, the effect sizes reported by Kraft et al. may be a better benchmark for interpreting the effect sizes in the current study than the effect sizes reported by Harris and Sass (2007, 2009).

From the current study's findings, we conclude that evidence about the positive effect of NBCT instructional leaders on teacher and student outcomes is encouraging. Although we did not find a statistically significant impact of NBCTs on mentored teachers' instructional practice, the effect sizes suggest that mentorship provided by NBCTs was associated with positive and meaningful improvements in the novice teachers' ability to provide Emotional Support and Classroom Organization. The study's results also suggest a positive relationship between NBCT mentorship and student achievement. Combined, our findings suggest that schools and districts investing or planning to invest in the National Board teacher certification may benefit from deploying NBCTs in roles that will position them to provide instructional support to novice teachers and potentially other teachers.

Limitations and Directions for Future Research

It is important to consider the current study's limitations in interpreting findings. One limitation is the study's design, which is a quasi-experiment. Although this design is a good proxy for experimental design, it does not account for unobserved differences between participants. For example, mentors could have asked to be paired with novice teachers from a certain school or teacher preparation program. Novice teachers could have differed in their level of motivation to participate in BTSA or engage with mentors. The study did not account for these possible differences, which could have biased results. Random assignment of novice teachers to mentors and random assignment of novice teachers to student rosters would have allowed to control for unobserved differences between participants. However, random assignment was not feasible within the BTSA program constraints.

Furthermore, while the study detected differences in the effect of NBCT mentors and non-NBCT mentors, we did not directly observe the supports that mentors provided to novice teachers. Therefore, we do not have knowledge about the difference in supports provided by NBCT and non-NBCT mentors. These differences could have been associated with the National Board Certification. These differences also could have also been associated with other factors, perhaps collaboration opportunities for NBCTs in San Francisco. Not knowing what drove the contrast between NBCT and non-NBCT mentors prevents us from confidently attributing the study's findings to the certification status of mentors.

Another substantial limitation of our study is the sample sizes. Because of the complications with data collection outside of the researchers' control (e.g., teachers' willingness to participate in observations, availability of baseline data), only 54% of all observed teachers could be included in the teacher sample to examine classroom practice. The sample did not provide sufficient power to estimate differences between the groups of novice teachers mentored by NBCTs and non-NBCTs at a statistically significant level. Using statistics derived from our sample, we estimate that the sample size for the domain of Emotional Support, for example, would need to be 136 teachers with 68 teachers in the intervention condition.[13] For comparison, our sample was 60 teachers with 12 in the intervention condition.

The issue of sample size was further exacerbated in the student achievement analysis, where we were only able to examine outcomes of students taught by three teachers mentored by

[13] We followed the procedure for sample size calculation described by Kadam and Bhalerao (2010). Because the standard deviation of the population is unknown, we used the postregression predicted values to estimate the standard deviation.

NBCTs. Even though the findings of the relationship between NBCT mentorship and student achievement are significant at the *p* value of .05 and have a meaningful effect size of 0.18 SDs, the generalizability of this finding is limited. With only three teachers teaching students in the intervention group, there is a possibility that characteristics of these teachers are associated with the study's findings, and not the fact that they were mentored by NBCTs. These characteristics could be, for example, the teachers' preparation program or the support they received in addition to NBCT mentor, perhaps from their colleagues or principal.

Study limitations notwithstanding, this study is the first attempt to rigorously examine the direct relationship between NBCTs in instructional leadership roles and teacher and student outcomes. We believe that the results are sufficiently compelling to warrant additional research using a more rigorous study design in which teachers are randomly assigned to NBCT and non-NBCT mentors, and in which mentored teachers are randomly assigned to student rosters. Any future research will also benefit from using larger samples, which can be estimated using statistics reported in the current study.

References

Association for Supervision and Curriculum Development (2014). *Teacher leadership: The what, why, and how of teachers as leaders.* A report on the Fall 2014 ASCD Whole Child Symposium. Alexandria, VA: Author. Retrieved from https://www.ascd.org/ASCD/pdf/siteASCD/wholechild/fall2014wcsreport.pdf

Backes, B., & Hansen, M. (2018). The impact of Teach For America on non-test academic outcomes. *Education Finance and Policy*, *13*(2), 168–193.

Belson, S., & Husted, T. (2015). Impact of National Board for Professional Teaching Standards Certification on student achievement. *Education Policy Analysis Archives, 23*(91), 1–21.

Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, *111*(3), 430–454.

Cannata, M., McCory, R., Sykes, G., Anagnostopoulos, D., & Frank, K. (2010). Exploring the influence of National Board Certified Teachers in their schools and beyond. *Educational Administration Quarterly, 46*(4), 463–490.

Cantrell, S., Fullerton, J., Kane, T. J., & Staiger, D. O. (2008). *National Board Certification and teacher effectiveness: Evidence from a random assignment experiment* (NBER Working Paper Series). Cambridge, MA: National Bureau of Economic Research.

Carver-Thomas, D., & Darling-Hammond, L. (2017). *Teacher turnover: Why it matters and what we should do about it*. Palo Alto, CA: Learning Policy Institute. Retrieved from https://learningpolicyinstitute.org/sites/default/files/product-files/Teacher_Turnover_REPORT.pdf

Cavalluzzo, L. (2004). *Is National Board Certification an effective signal of teacher quality?* Alexandria, VA: The CNA Corporation. Retrieved from https://files.eric.ed.gov/fulltext/ED485515.pdf

Center for Strengthening the Teaching Profession. (2016b). *Teacher leadership* [web page]. Retrieved from http://cstp-wa.org/teacher-leadership/

Cohen, B., & Fuller, E. (2006). *Effects of mentoring and induction on beginning teacher retention.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Cowan, J., & Goldhaber, D. (2016). National Board Certification and teacher effectiveness: Evidence from Washington state. *Journal of Research on Educational Effectiveness, 9*(3), 233–258.

Evertson, M., & Smithey, M. W. (2000). Mentoring effects on protégés' classroom practice: An experimental field study. *Journal of Educational Research, 93*(5), 294–304.

Fuller, E. (2003). *Beginning teacher retention rates for TxBESS and Non-TxBESS teachers.* Unpublished paper, State Board for Educator Certification, Texas.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, E., Grider, M., & Jacobus, M. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study*. Washington, DC: U.S. Department of Education.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National Board Certification as a signal of effective teaching. *Review of Economics and Statistics, 89*(1), 134–150.

Harris, D. N., & Sass, T. R. (2007). The effects of NBPTS-certified teachers on student achievement. *Working paper.* Retrieved from https://files.eric.ed.gov/fulltext/ED509659.pdf

Harris, D. N., & Sass, T. R. (2009). The effects of NBPTS-certified teachers on student achievement. *Journal of Policy Analysis and Management, 28*(1), 55–80.

Hanushek, E. A., Woessmann, L., & Peterson, P. E. (2012). Is the US catching up? *Education Next*, *12*(4).

Henry, G., Fortner, K., & Bastian, K. (2012). The effects of experience and attrition for novice high-school science and mathematics teachers. *Science, 335*(2), 1118–1121. Retrieved from https://pdfs.semanticscholar.org/8442/d3ff6abceb3f56994fde0c3a5375e598af7e.pdf?_ga=2.219690585.1350962984.1556292412-1032698858.1548363548

Ingersoll, R., & Smith, T. M. (2004). Do teacher induction and mentoring matter? *NASSP Bulletin, 88*(638), 28–40.

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

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, *1*(4), 85–108.

Johnson, L., Goldrick, L., & Lasagna, M. (2010). *New teacher excellence: The impact of state policy on induction program implementation* [policy brief]. Santa Cruz, CA: New Teacher Center. Retrieved from http://newteachercenter.org/wp-content/uploads/NTC_Policy_Brief-NewTeacherExcellence.pdf

Kadam, P., & Bhalerao, S. (2010). Sample size calculation. *International journal of Ayurveda research*, 1(1), 55–57.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from https://eric.ed.gov/?id=ED540960

Kapadia, K., Coca, C., & Easton, J. Q. (2007). *Keeping new teachers: A first look at the influences of induction in the Chicago Public Schools*. Chicago, IL: Consortium on Chicago School Research.

Kini, T., & Podolsky, A. (2016). *Does teaching experience increase teacher effectiveness? A review of the research*. Palo Alto, CA: Learning Policy Institute. Retrieved from https://learningpolicyinstitute.org/sites/default/files/product-files/Teaching_Experience_Report_June_2016.pdf

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588.

Levenson, M. (2014). *Pathways to teacher leadership: Emerging models, changing roles*. Harvard Education Press: Cambridge, MA.

Loeb, H., Elfers, A., Plecki, M., Ford, B., & Knapp, M. (2006). *National Board Certified Teachers in Washington state: Impact on professional practice and leadership opportunities*. Seattle, WA: University of Washington, Center for Strengthening the Teaching Profession. Retrieved from https://www.education.uw.edu/ctp/sites/default/files/ctpmail/PDFs/NBCTinWA.pdf

Manzeske, D., Park, S., Feng, L., Borman, T., Gnedko-Berry, N., West, B., & Deng, E. (2017). *Effects of National Board Certified Teachers on student achievement and behavioral outcomes: Studies conducted in two states* [working paper]. Washington, DC: American Institutes for Research. Retrieved from https://files.eric.ed.gov/fulltext/ED572969.pdf

Mihaly, K., Master, B., & Yoon, C. (2015). *Examining the early impacts of the Leading Educators Fellowship on student achievement and teacher retention*. Santa Monica, CA: Rand Corporation.

Mitchell, D., Scott-Hendrick, L., Parrish, T., Crowley, J., Boyns, D., & Mitchell, T. D. (with Batie, M., Grindal, M., Hale, T., Knox, M., Naylor, N. J., Romero, L.). (2007). *California Beginning Teacher Support and Assessment and Intern Alternative Certification Evaluation Study technical report*. Riverside, CA: University of California Riverside. Retrieved from http://www.ctc.ca.gov/reports/BTSA-Intern-Technical-Report-23-Oct-2007.pdf

National Board (2016). *What teachers should know and be able to do*. Arlington, VA: Author. Retrieved from http://accomplishedteacher.org/wp-content/uploads/2016/12/NBPTS-What-Teachers-Should-Know-and-Be-Able-to-Do-.pdf

National Board (2018). 2018 state rankings by total number of National Board Certified Teachers. Retrieved from https://www.nbpts.org/wp-content/uploads/StateRankings_All_NBCTs.pdf

O'Dwyer, L. M., and Parker, C. E. (2014). *A primer for analyzing nested data: Multilevel modeling in SPSS using an example from a REL study* (REL 2015–046). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2015046.pdf

Pianta, R. C., Hamre, B. K., & Mintz, S. (2012a). *Classroom Assessment Scoring System manual: Secondary*. Charlottesville, VA: Teachstone.

Pianta, R. C., Hamre, B. K., Mintz, S. (2012b). *Classroom Assessment Scoring System manual: Upper elementary.* Charlottesville, VA: Teachstone.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System manual: K–3*. Baltimore, MD: Paul H. Brookes Publishing Co.

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

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal, 50*(1), 4–36.

Ryan, S. (1999, April). *Principals and teachers leading together.* Paper presented at the annual meeting of the American Educational Research Organization. Montreal, Quebec, Canada. Retrieved from https://eric.ed.gov/?id=ED440457

San Francisco Unified School District (SFUSD). (2015). Accreditation: Draft induction program standards and preconditions. Retrieved from https://www.ctc.ca.gov/docs/default-source/commission/agendas/2015-12/2015-12-2g-pdf.pdf.

San Francisco Unified School District (2018). BTSA/Induction program. Available at http://www.sfusdbtsa.org.

Sato, M., Hyler, M., & Monte-Sano, C. (2014). Learning to lead with purpose: National Board Certification and teacher leadership. *International Journal of Teacher Leadership, 5*(1), 1–23.

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

Silva, D., Gimbert, B., & Nolan, J. (2000). Sliding the doors: Locking and unlocking possibilities for teacher leadership. *Teachers College Record, 102*(4), 779–804.

Stanulis, R., & Floden, E. (2009). Intensive mentoring as a way to help beginning teachers develop balanced instruction. *Journal of Teacher Education, 60*(2), 112–122.

Strong, M. (2005). *Mentoring new teachers to increase retention: A look at the research*. Santa Cruz, CA: New Teacher Center.

Strong, M. (2006). *Does teacher support affect student achievement?* Santa Cruz, CA: New Teacher Center.

Strong, M. (2009). *Effective teacher induction and mentoring: Assessing the evidence*. New York, NY: Teachers College Press.

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2016). *A coming crisis in teaching? Teacher supply, demand, and shortages in the U.S.* Palo Alto, CA: Learning Policy Institute.

Thompson, M., Paek, P., Goe, L., & Ponte, E. (2004). *Study of the impact of the California Formative Assessment and Support System for Teachers Report 2: Relationship of BTSA/CFASST engagement and teacher practices* (ETS-RR-04-31). Washington, DC: Educational Testing Service.

Tushnet, N., Briggs, D., Elliott, J., Esch, C., Haviland, D., Humphrey, D. C., et al. (2002). *Final report of the independent evaluation of the Beginning Teacher Support and Assessment Program (BTSA).* San Francisco, CA: WestEd.

What Works Clearinghouse. (2017). *Standards handbook* (version 4.0). Washington, DC: Author. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.

York-Barr, J., & Duke, K. (2004). What do we know about teacher leadership? Findings from two decades of scholarship. *Review of Educational Research*, *74*(3), 255–316.

Yuan, K. (2015). A value-added study of teacher spillover effects across four core subjects in middle schools. *Education Policy Analysis Archives*, *23*, 38.

Appendix A. Characteristics of Original and Analytic Samples for the Analysis of Teacher Classroom Practice

We compared average CLASS scores and other characteristics of teachers removed from the analytic sample and teacher kept in the analytic sample. We examined whether any differences were statistically significant between the two groups. We examined differences using Hedges' *g* to compute standardized mean differences for continuous variables and the Cox index for binary variables. The differences between groups were smaller than 0.25 standard deviations (SDs) on most average CLASS scores and characteristics, except for the baseline CLASS score in the Classroom Management domain and full-time employment status. All differences were not statistically significant, suggesting that the two samples were similar. Results are shown in Table A1.

Table A1. Average CLASS Score and Characteristics of the Original Sample and Analytic Sample[a]

Covariates	Analytic Sample		Removed Sample[b]		Standardized Mean Difference
	Intervention (n = 12)	Comparison (n = 48)	Intervention (n = 16)	Comparison (n = 35)	
CLASS Scores					
Baseline CLASS score in the Emotional Support domain (standardized)	0.79	0.90	1.74	0.92	0.10
Baseline CLASS score in the Classroom Management domain (standardized)	0.61	0.38	2.01	0.83	0.34
Baseline CLASS score in the Instructional Support domain (standardized)	-0.55	-0.21	1.16	-0.10	0.18
Outcome data on CLASS score in the Emotional Support domain (standardized)	0.92	0.84	0.69	0.53	-0.18
Outcome data on CLASS score in the Classroom Management domain (standardized)	0.69	0.21	-0.25	0.13	-0.18
Outcome data on CLASS score in the Instructional Support domain (standardized)	-0.41	-0.40	-0.61	-0.24	0.00
Teacher Covariates					
Racial/ethnic minority	42%	38%	50%	38%	0.08
Full-time employment status	83%	88%	94%	94%	0.50
Female gender	92%	63%	69%	72%	0.07
Grades taught (K–3)	33%	21%	31%	29%	0.19
Grades taught (secondary)	25%	48%	25%	40%	-0.20

Covariates	Analytic Sample		Removed Sample[b]		Standardized Mean Difference
	Intervention (n = 12)	Comparison (n = 48)	Intervention (n = 16)	Comparison (n = 35)	
Grades taught (upper elementary)	42%	31%	44%	31%	0.05
School Covariates					
Total enrollment	427.63	796.44	782.75	679.89	-0.02
EL students	30%	28%	30%	30%	0.11
FRPL-eligible students	56%	61%	70%	60%	0.19
Students with disability	14%	13%	15%	13%	0.02
Racial/ethnic minority students	85%	88%	92%	90%	0.23

[a] No statistically significant differences were found between teachers removed from the analytic sample and teachers kept in the analytic sample.

[b] Numbers of cases with available data varied by covariate.

Appendix B. Standardized Mean Differences Between Intervention and Comparison Groups at Baseline

Tables B1 and B2 present standardized mean differences between the intervention and comparison groups (i.e., teachers, students) at baseline.

Table B1. Standardized Mean Differences Between Intervention and Comparison Teachers at Baseline

Baseline Covariates	Intervention Group Mean (Standard Deviation)	Comparison Group Mean (Standard Deviation)	Standardized Mean Difference
CLASS Scores			
Baseline CLASS score in the Emotional Support domain (standardized)	0.79 (1.04)	0.90 (1.34)	-0.08
Baseline CLASS score in the Classroom Management domain (standardized)	0.61 (0.93)	0.38 (1.63)	0.15
Baseline CLASS score in the Instructional Support domain (standardized)	-0.55 (1.86)	-0.21 (1.51)	-0.21
Baseline CLASS global score (standardized)	0.35 (1.27)	0.35 (1.27)	-0.06
Teacher Characteristics			
Racial/ethnic minority	42% (0.51)	38% (0.49)	0.10
Full-time employment status	83% (0.39)	88% (0.33)	-0.20
Female gender	92% (0.29)	63% (0.49)	1.13
Grades taught (K–3)	33% (0.49)	21% (0.41)	0.38
Grades taught (secondary)	25% (0.45)	48% (0.50)	-0.61
Grades taught (upper elementary)	42% (0.51)	31% (0.47)	0.27

Baseline Covariates	Intervention Group Mean (Standard Deviation)	Comparison Group Mean (Standard Deviation)	Standardized Mean Difference
School Characteristics			
Total enrollment	427.63 (269.16)	796.44 (711.62)	-0.56
EL students	30% (0.22)	28% (0.16)	0.13
FRPL-eligible students	56% (0.22)	61% (0.15)	-0.30
Students with disability	14% (0.05)	13% (0.05)	0.29
Racial/ethnic minority students	85% (0.15)	88% (0.13)	-0.25

Note. EL = English learner; FRPL = free- and reduced-price lunch

Table B2. Standardized Mean Differences Between Intervention and Comparison Students at Baseline

Baseline Covariates	Intervention Group Mean (Standard Deviation)	Comparison Group Mean (Standard Deviation)	Standardized Mean Difference
Prior score in the measure of general achievement (standardized)	0.64 (0.77)	0.59 (0.74)	0.04
Female gender	47% (0.50)	58% (0.50)	-0.24
Racial/ethnic minority	82% (0.39)	96% (0.19)	-1.13
EL status	9% (0.29)	9% (0.29)	0.48
SPED status	12% (0.32)	4% (0.19)	0.22

Note. ELA = English language arts; EL = English learner; SPED = special education

Appendix C. CLASS Data

CLASS domains and dimensions are described in Table C1 (descriptions are from the CLASS manual).[14] The dimensions within each domain are represented as they were used in the current study.

Table C1. Description of CLASS Domains and Dimensions Shared Across All Three Levels

CLASS Domain	Description
Emotional Support	Teacher's ability to support social and emotional functioning in the classroom
Classroom Organization	The classroom processes aimed at organizing and managing students' behavior, time, and attention
Instructional Support	The ways teachers implement their curriculum to support cognitive and language development
CLASS Dimension	**Description**
Emotional Support Domain	
Positive Climate	The emotional connection, respect, and enjoyment demonstrated between teachers and students and among students
Negative Climate	The level of expressed negativity such as anger, hostility, or aggression exhibited by teachers and/or students
Teacher Sensitivity	Teachers' awareness of and responsivity to students' academic and emotional concerns
Regard for Student/Adolescent Perspectives[15]	The degree to which teachers meet the social and developmental needs and goals of students for decision-making, having their opinions valued, and meaningful interactions with peers
Classroom Organization Domain	
Behavior Management	How effectively teachers monitor, prevent, and redirect behavior
Productivity	How well the classroom runs (based on routines and organization of activities and directions) so that time on learning activities can be maximized

[14] Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System manual: K–3.* Baltimore, MD: Paul H. Brookes Publishing Co.

[15] At the K–3 and Upper upper elementary levels, this dimension is Regard for Student Perspectives. At the secondary level, it is Regard for Adolescent Perspectives. We have received permission from Teachstone to treat these dimensions as identical.

Instructional Learning Formats	How teachers facilitate activities and provide interesting materials to engage students and maximize learning opportunities
Instructional Support Domain	
Quality of Feedback	How teachers extend students' learning through responses to students' ideas, comments, and work

To pool all teacher scores across grade-levels, the CLASS scores were standardized using the means reported in the CLASS manuals. These means are shown in Tables C2–C4 (by grade-level).[16] Scores for each domain were calculated by averaging standardized scores between one and five associated dimensions.

Table C2. CLASS Score Descriptives: Grades K–3, Raw Sample Mean & Standard Deviation From Manual

	From CLASS Manual	
	Mean	Standard Deviation
Positive Climate	5.16	0.84
Negative Climate	6.43	0.68
Teacher Sensitivity	4.66	0.93
Regard for Student Perspectives	4.38	1.00
Behavior Management	5.08	0.90
Productivity	4.71	0.84
Instructional Learning Formats	4.10	0.96
Quality of Feedback	2.28	0.84

[16] The CLASS Manual means for K–3 and secondary are the pooled means and standard deviations from multiple datasetsdatasetsdata sets.

Table C3. CLASS Score Descriptives: Upper Elementary (Grades 4–6), Raw Sample Mean & Standard Deviation From Manual

	From CLASS Manual	
	Mean	Standard Deviation
Positive Climate	4.68	0.61
Negative Climate	6.68	0.35
Teacher Sensitivity	4.26	0.55
Regard for Student Perspectives	3.29	0.60
Behavior Management	6.01	0.58
Productivity	5.91	0.46
Instructional Learning Formats	4.36	0.52
Quality of Feedback	3.76	0.57

Table C4. CLASS Score Descriptives: Secondary (Grades 7–12), Raw Sample Mean & Standard Deviation From Manual

	From CLASS Manual	
	Mean	Standard Deviation
Positive Climate	3.97	0.69
Negative Climate	6.43	0.50
Teacher Sensitivity	3.92	0.64
Regard for Student Perspectives	2.88	0.71
Behavior Management	5.51	0.69
Productivity	5.42	0.69
Instructional Learning Formats	3.84	0.67
Quality of Feedback	3.19	0.67

Appendix D. Statistical Approach to Estimate Differences in Teacher Classroom Practice

To estimate differences in teacher classroom practice, we used a hierarchical linear modeling design where teachers were nested in schools. Analyses were conducted separately by domains (i.e., Emotional Support, Classroom Organization, Instructional Support, and global score across the three domains). The statistical model for the analysis is as follows:

$$Y_{js} = \beta_0 + \beta_1 Intervention_{js} + \beta_2 PreCLASS_{js} + \beta_3 TeacherDemo_{js} + \beta_4 SchoolDemo_s + \beta_5 Cohort_j + u_s + v_{js}$$

Formally, Y_{js} represents the standardized CLASS score in one given domain of teacher *j* in school *s*, $Intervention_{js}$ is an indicator for whether the teacher was in the intervention group (coded 0 if the teacher was in the comparison group and 1 if the teacher was in the intervention group), and $PreCLASS_{js}$ is the teacher's baseline standardized CLASS score in the same domain. $TeacherDemo_{js}$ is a vector of teacher demographic characteristics—gender, race/ethnicity, full-time employment status, and grade levels taught (i.e., K–3, upper elementary, or secondary). $SchoolDemo_s$ is a vector of school-level characteristics—total enrollment and the percentage of EL, disabled, female, and FRPL-eligible students at each school. $Cohort_j$ is a dummy variable for cohort (coded 1 if the teacher was observed in 2015-16 and 0 if the teacher was observed in 2016-17). Random effects are included to account for school and teacher effects by adding a random error term for each school (u_s) and teacher (v_{js}).

The coefficients (β) represent the relationship between covariates and the outcome. The coefficient of interest, β_1, represents the difference in average outcomes between intervention and comparison teachers. This coefficient represents the relationship between being mentored by NBCTs and the classroom practice of novice teachers.

Appendix E. Statistical Approach to Estimate Differences in Student Achievement

To estimate differences in student achievement, we used regression analyses. Analyses were conducted using the combined ELA and mathematics samples. The statistical model for the analysis is as follows:

$$Y_{ijs} = \beta_0 + \beta_1 Intervention_j + \beta_2 PreAchievement_i + \beta_3 PreCLASS_j + \beta_4 StudentDemo_i + \beta_5 Grade_i + u_i$$

Formally, Y_i represents general achievement, measured by the standardized score in ELA or mathematics assessments of student *i* taught by teacher *j* in school *s*, $Intervention_j$ is an indicator for whether the teacher was in the intervention group (coded 0 if the teacher was in the comparison group and 1 if the teacher was in the intervention group), $PreAchievement_i$ is the student's scores in the same assessment in the prior school year, and $PreCLASS_j$ is the teacher's baseline standardized CLASS score in one domain (i.e., Emotional Support, Classroom Organization, Instructional Support) or global score. $StudentDemo_i$ is a vector of student background characteristics (e.g., gender, race/ethnicity, EL status, SPED status). $Grade_i$ is a fixed effect to ensure students were compared within grades. u_i is student-level residual error, assumed to be independently and identically distributed.

The coefficients (β) represent the relationship between covariates and the outcome. The coefficient of interest, β_1, represents the difference in average outcomes between intervention and comparison students. This coefficient represents the relationship between being taught by NBCT-mentored teachers and student achievement.

Appendix F. Regression Results

Estimated Differences in Teacher Classroom Practice

Table F1. Estimated Differences in Teacher Classroom Practice Between Intervention and Comparison Teachers

Term	Outcome 1 (Global Score)	Outcome 2 (Emotional Support)	Outcome 3 (Classroom Organization)	Outcome 4 (Instructional Support)
Constant	1.231 (1.467)	1.510 (1.550)	2.027 (1.632)	0.184 (1.992)
Intervention	0.266 (0.379)	0.367 (0.407)	0.412 (0.426)	-0.102 (0.479)
Baseline standardized CLASS score	0.288 (0.126)	0.194 (0.143)	0.426*** (0.114)	0.102 (0.122)
Racial/ethnic minority	0.135 (0.300)	-0.202 (0.328)	0.0349 (0.338)	0.510 (0.373)
Full-time employment status	-0.422 (0.451)	-0.0515 (0.485)	-0.367 (0.511)	-0.771 (0.554)
Female gender	-0.251 (0.331)	-0.427 (0.356)	-0.477 (0.370)	0.080 (0.414)
Grade level taught (K–3)	0.220 (0.448)	-0.670 (0.491)	0.235 (0.502)	0.977 (0.589)
Grade level taught (upper elementary)	0.565 (0.475)	0.354 (0.508)	0.410 (0.545)	0.916 (0.580)
Total enrollment	-0.000 (0.000)	-0.000 (0.000)	-0.001 (0.000)	0.000 (0.001)
Percent EL students	0.673 (1.343)	0.791 (1.431)	-0.124 (1.507)	1.091 (1.826)
Percent FRPL-eligible students	-2.118 (2.330)	-1.599 (2.450)	-2.359 (2.558)	-0.445 (3.228)
Percent students with disability	-6.288 (3.916)	-4.016 (4.092)	-8.756* (4.296)	-4.970 (5.626)
Percent racial/ethnic minority students	1.485 (2.492)	1.248 (2.600)	1.709 (2.716)	-0.169 (3.468)
Cohort 2	-0.284 (0.346)	-0.357 (0.378)	-0.0691 (0.383)	-0.210 (0.464)

Note. * indicates significant difference at the $p \leq 0.05$ level; *** indicates significant difference at the $p \leq 0.001$ level. The analyses included 12 intervention teachers and 48 comparison teachers.

Estimated Differences in Student General Achievement

Table F2. Estimated Differences in Student General Achievement Between Intervention and Matched Comparison Students Using Different Covariates

Term	Main Model (Global Baseline CLASS Score as Covariate)
Constant	0.166 (0.125)
Intervention	0.134* (0.069)
Baseline standardized test score	0.819*** (0.0412)
Female gender	0.0308 (0.0518)
Racial/ethnic minority	0.00610 (0.100)
EL status	0.00697 (0.109)
SPED status	-0.135 (0.0890)
Student grade (sixth grade)	-0.286*** (0.0842)
Student grade (seventh grade)	-0.296** (0.101)
Teacher average baseline CLASS score	0.0728 (0.0581)

Note. ES = Emotional Support; CO = Classroom Organization; IS = Instructional Support. All models included student prior achievement and other student characteristics. Baseline CLASS scores in specific domains included in regression models are presented in parentheses: Model 1 included teacher's average global baseline CLASS score (main model); model 2 included teacher's average baseline CLASS score in ES domain; model 3 included teacher's average baseline CLASS score in CO domain; model 4 included teacher's average baseline CLASS score in IS domain. Estimates are shown in standard deviations. Standard errors are presented in parentheses. * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Appendix G. Results of Sensitivity Analysis for Student Achievement

Table G1. Estimated Differences in Student General Achievement Between Intervention and Matched Comparison Students Using Different Covariates

Included Additional Covariates	Estimates	*p* Value	Effect Size
Teacher's average baseline CLASS score in Emotional Support domain	-0.024 (0.120)	.844	-0.031
Teacher's average baseline CLASS score in Classroom Organization domain	0.158** (0.055)	.005	0.210**
Teacher's average baseline CLASS score in Instructional Support domain	0.238** (0.072)	.001	0.316**

Note. Analyses included 116 intervention students and 57 comparison students. All models included student prior achievement and other student characteristics. Estimates are shown in standard deviations. Standard errors are presented in parentheses. * $p \leq .05$; ** $p \leq .01$

Table G2. Estimated Differences in Student General Achievement Between Intervention and Matched Comparison Students Using Different Covariates

Term	Model 1 (Baseline CLASS Score in ES as Covariate)	Model 2 (Baseline CLASS Score in CO as Covariate)	Model 3 (Baseline CLASS Score in IS as Covariate)
Constant	0.200 (0.124)	0.151 (0.123)	0.125 (0.131)
Intervention	-0.024 (0.120)	0.158** (0.0551)	0.238** (0.0720)
Baseline standardized test score	0.824*** (0.0427)	0.820*** (0.0408)	0.813*** (0.0421)
Female gender	0.0271 (0.0519)	0.0329 (0.0512)	0.0383 (0.0531)
Racial/ethnic minority	0.00214 (0.0967)	-0.00149 (0.1000)	0.0244 (0.101)
EL status	-0.0292 (0.116)	0.0135 (0.109)	0.0118 (0.110)
SPED status	-0.120 (0.0950)	-0.142 (0.0883)	-0.151 (0.0896)

Term	Model 1 (Baseline CLASS Score in ES as Covariate)	Model 2 (Baseline CLASS Score in CO as Covariate)	Model 3 (Baseline CLASS Score in IS as Covariate)
Student grade (sixth grade)	-0.557** (0.182)	-0.248** (0.0866)	-0.309** (0.0948)
Student grade (seventh grade)	-0.541** (0.187)	-0.269** (0.0955)	-0.267** (0.0975)
Teacher baseline average CLASS score	0.205 (0.108)	0.0625* (0.0312)	-0.0452 (0.0536)

Note. ES = Emotional Support; CO = Classroom Organization; IS = Instructional Support. All models included student prior achievement and other student characteristics. Baseline CLASS scores in specific domains included in regression models are presented in parentheses: Model 1 included teacher's average global baseline CLASS score (main model); model 2 included teacher's average baseline CLASS score in ES domain; model 3 included teacher's average baseline CLASS score in CO domain; model 4 included teacher's average baseline CLASS score in IS domain. Estimates are shown in standard deviations. Standard errors are presented in parentheses. * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

AIR®

AMERICAN INSTITUTES FOR RESEARCH®

Established in 1946, the American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research on important social issues and delivers technical assistance, both domestically and internationally, in the areas of education, health, and workforce productivity.

MAKING RESEARCH **RELEVANT**

AMERICAN INSTITUTES FOR RESEARCH
1000 Thomas Jefferson Street NW
Washington, DC 20007-3835 | 202.403.5000
www.air.org

LOCATIONS

Domestic: Washington, DC (HQ) | Monterey, Sacramento, and San Mateo, CA | Atlanta, GA | Honolulu, HI | Chicago and Naperville, IL Indianapolis, IN | Metairie, LA | Waltham, MA | Frederick and Rockville, MD | Chapel Hill, NC | New York, NY | Columbus, OH | Cayce, SC Austin, TX | Arlington and Reston, VA

International: El Salvador | Ethiopia | Haiti | Honduras | Zambia

8257_05/19