IES Institute of Education Sciences

Appendix

Study of Teacher Coaching Based on Classroom Videos: Impacts on Student Achievement and Teachers' Practices

NCEE-2022-006a
U.S. Department of Education

A Publication of the National Center for Education Evaluation at IES

The Institute of Education Sciences (IES) is the independent, non-partisan statistics, research, and evaluation arm of the U.S. Department of Education. The IES mission is to provide scientific evidence on which to ground education practice and policy and to share this information in formats that are useful and accessible to educators, parents, policymakers, researchers, and the public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other IES product or report, we would like to hear from you. Please direct your comments to ncee.feedback@ed.gov.

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-16-C-0021 by Mathematica. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

June 2022

Clark, M., Max, J., James-Burdumy, S., Robles, S., McCullough, M., Burkander, P., and Malick, S. (2022). Study of Teacher Coaching Based on Classroom Videos: Impacts on Student Achievement and Teachers' Practices: Appendix (NCEE 2022-006a). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee.

This report is available on the Institute of Education Sciences website at http://ies.ed.gov/ncee.

IES Institute of Education Sciences

Study of Teacher Coaching Based on Classroom Videos: Impacts on Student Achievement and Teachers' Practices

June 2022

Melissa Clark
Jeffrey Max
Susanne James-Burdumy
Silvia Robles
Moira McCullough
Paul Burkander
Steven Malick
Mathematica

NCEE 2022-006a
U.S. DEPARTMENT OF EDUCATION

CONTENTS

EXHIBITS

INTRODUCTION

Helping teachers become more effective in the classroom is a high priority for educators and policymakers. A growing body of evidence suggests that individualized coaching focused on general teaching practices can improve teachers' instruction and student achievement. However, little is known about the benefits of specific approaches to coaching, including who is doing the coaching, how coaches observe teachers' instruction, and how or how often coaches provide feedback to teachers. This study examined one promising strategy for individualized coaching: professional coaches—rather than district or school staff—providing feedback to teachers based on videos of their instruction. Feedback based on videos gives teachers the opportunity to observe and reflect on their own teaching and allows coaches to show teachers specific moments from their teaching when providing feedback. For this study, 107 elementary schools were randomly divided into three groups: one that received fewer highly structured cycles of focused professional coaching during a single school year (five cycles), one that received more (eight cycles), and one that continued with its usual strategies for supporting teachers. The study compared teachers' experiences and student achievement across the three groups to determine the effectiveness of the two versions of the coaching. This document provides additional details on the coaching provided for the study, the approach to carrying out the study, and the findings presented in the report.

APPENDIX A. THE STUDY'S VIDEO-BASED COACHING FOR TEACHERS

A.1 Overview of the study's coaching

Teachstone provided the study's coaching to teachers using its MyTeachingPartner program. The program aims to provide specific and actionable feedback to improve teachers' practices in ways that are intended to improve student achievement. Coaches watch videos of teachers' instruction and provide feedback remotely via videoconferences (or, when videoconferences are not feasible, via phone calls).

Teachstone's coaching program focuses on general teaching practices rather than practices specific to a particular curriculum or subject area. Specifically, the program focuses on teaching practices from the Classroom Assessment Scoring System (CLASS) observation rubric. The CLASS is based on the theory that interactions between teachers and students are critical to students' development, with effective teachers actively engaging students and creating environments that are conducive to learning.

The CLASS organizes teaching practices into three broad domains: (1) classroom management, (2) building students' understanding, and (3) building supportive relationships with students.[1] The classroom management domain covers teachers' practices to organize and manage students' behavior, establish efficient classroom routines, and avoid a negative classroom climate. The building students' understanding domain covers how teachers structure lessons and activities to develop students' understanding of the content, provide opportunities for higher-level thinking, and engage students in content-focused discussions. Finally, the building supportive relationships with students domain covers practices that build a positive learning environment, respond to students' social-emotional needs, and connect the content to students' lives. Each domain is made up of more detailed aspects of teaching called dimensions. For example, within the classroom management domain, the dimension of behavior management describes how teachers set clear expectations for student behavior and anticipate and redirect problem behavior. Finally, each dimension is made up of a set of specific teaching practices, which are referred to as "indicators" in the CLASS rubric (Exhibit A.1).

Exhibit A.1. Teaching practices (indicators) covered by the CLASS, by domain and dimension

Classroom management	
Behavior management	
Clear expectations	Teacher is clear about expectations for student behavior
Proactive	Teacher anticipates problem behavior
Effective redirection of misbehavior	Teacher redirects or solves problem behavior, or encourages students to redirect or solve problem behavior
Student behavior	Students readily cooperate with the teacher
Productivity	
Maximizing learning time	Teacher minimizes disruptions to learning
Routines	Teacher sets up clear routines
Transitions	Teacher helps students switch tasks quickly
Preparation	Teacher has materials and lessons ready

Negative climate	
Negative affect	Teacher or students display anger, irritability, or a negative attitude
Punitive control	Teacher uses yelling, threats, and punishment to control the class
Disrespect	Teacher or students tease, bully, or use discriminatory or disrespectful behavior towards others
Building students' understanding	
Instructional learning formats	
Learning targets/organization	Teacher presents information in a clear and organized way, with clear learning targets
Variety of modalities, strategies, and materials	Teacher uses multiple approaches to teach lesson content
Active facilitation	Teacher promotes student involvement by asking questions about and demonstrating interest in student work and ideas
Effective engagement	Teacher fosters student engagement in learning
Content understanding	
Depth of understanding	Teacher helps students gain a deeper understanding of content and helps them see how facts link to broader concepts or ideas
Communication of concepts and procedures	Teacher explains content clearly with multiple examples
Background knowledge and misconceptions	Teacher connects new ideas to students' prior knowledge
Transmission of content knowledge and procedures	Teacher provides clear and accurate definitions and clarifications
Opportunity for practice of procedures and skills	Teacher provides opportunities for students to practice skills with support and on their own
Analysis and inquiry	
Facilitation of higher-order thinking	Teacher provides opportunities to use higher-order thinking skills
Opportunities for novel application	Teacher provides opportunities to apply skills in new settings
Metacognition	Teacher models the thinking process and helps students explain their thinking
Quality of feedback	
Feedback loops	Teacher asks a series of follow-up questions to extend students' thinking or encourages back and forth exchanges among students
Scaffolding	Teacher provides support when students struggle with a concept or has other students provide support
Building on student responses	Teacher expands on and clarifies students' responses or has students expand on or clarify each other's responses
Encouragement and affirmation	Teacher encourages, praises, and supports students' efforts or encourages students to encourage, praise, and support each other's efforts

Instructional dialogue	
Cumulative, content-driven exchanges	Teacher encourages content-focused discussions that build on each other over time
Distributed talk	Teacher leads classroom discussions that involve many students
Facilitation strategies	Teacher asks open-ended questions and actively listens to facilitate productive conversations among students
Building supportive relationships with students	
Positive climate	
Relationships	Teacher provides opportunities for close, positive social interactions with teacher and peers
Positive affect	Teacher displays a positive attitude, including smiling, laughing, and showing enthusiasm
Positive communications	Teacher makes positive comments and conveys positive expectations
Respect	Teacher teaches and models respectful behavior
Teacher sensitivity	
Awareness	Teacher notices how students are doing and anticipates difficulties
Responsiveness to academics and social/emotional cues	Teacher responds to student needs with reassurance, support, and understanding
Effectiveness in addressing problems	Teacher addresses students' problems and follows up as needed
Student comfort	Teacher fosters classroom where students feel comfortable participating, taking risks, and asking for help
Regard for student perspectives	
Flexibility and student focus	Teacher encourages students to share ideas and adapts lesson plans to follow students' leads
Connections to current life	Teacher shows how content is relevant to students' lives
Support for autonomy and leadership	Teacher provides opportunities for students to lead, make choices, and take on responsibilities
Meaningful peer interactions	Teacher provides meaningful tasks that students accomplish together as a group
Building student engagement (stand-alone dimension)	
Active engagement	Teacher fosters students' active engagement in classroom activities

A.2 Focus and structure of the coaching

This section describes the specific focus and structure of the coaching delivered for the study.

A.2.1 Focus of the coaching

Although the coaching focused on teachers' general teaching practices, coaches and teachers worked together to apply these practices to teachers' math and ELA instruction. Teachers in self-contained classrooms who taught both math and English language arts chose which of the two subjects to record and reflect on in each cycle. The teacher and coach worked together to select two CLASS dimensions to focus on in each coaching cycle.

Teachstone recommended that coaches cover a specific sequence of CLASS domains across the assigned set of cycles (Exhibit A.2). During the first cycle, the coach and the teacher built a rapport and discussed the teacher's goals. The sequence then began in the second cycle with a focus on dimensions related to classroom management and building supportive relationships with students. The second cycle also focused on one dimension in the domain of building students' understanding—instructional learning formats. This dimension addresses how teachers clearly communicate learning objectives, provide interesting materials in a variety of learning formats, and actively facilitate student involvement in activities and discussions. The second and third cycles were intended to help teachers lay a foundation for a supportive classroom climate with well-managed student behavior and well-organized instruction. Coaches focused on dimensions related to building students' understanding for the remainder of the year.

Although Teachstone recommended that coaches follow this sequence, they were not required to do so. Ultimately, coaches and teachers worked together to select the areas of focus for the coaching based on teachers' needs. However, most teachers (75 percent of those in the five-cycle group and 79 percent of those in the eight-cycle group) received coaching that followed the recommended progression of CLASS domains.[2]

Exhibit A.2. Recommended sequence of domains for the coaching program, by coaching group, 2018–2019 school year

Cycle number	Five-cycle coaching group		Eight-cycle coaching group	
	First focus area	Second focus area	First focus area	Second focus area
1	Getting to know you	Getting to know you	Getting to know you	Getting to know you
2	Building supportive relationships with students	Building students' understanding: Instructional learning formats	Building supportive relationships with students	Building students' understanding: Instructional learning formats
3	Classroom management	Building students' understanding	Classroom management	Building students' understanding
4	Building students' understanding	Building students' understanding	Classroom management or building students' understanding	Building students' understanding
5	Building students' understanding	Building students' understanding	Building students' understanding	Building students' understanding
6	n.a.	n.a.	Building students' understanding	Building students' understanding
7	n.a.	n.a.	Building students' understanding	Building students' understanding

Cycle number	Five-cycle coaching group		Eight-cycle coaching group	
	First focus area	Second focus area	First focus area	Second focus area
8	n.a.	n.a.	Building students' understanding	Building students' understanding

Source: Teachstone training materials.

n.a. = not applicable.

The coaching focused primarily on CLASS dimensions related to building students' understanding (Exhibit A.3). On average, consistent with the recommended sequence, teachers in the five-cycle group spent just under four cycles and teachers in the eight-cycle group spent just under seven cycles focused on dimensions related to building students' understanding. These dimensions included engaging students through clear, interesting lessons (instructional learning formats); building students' understanding of core academic content (content understanding); supporting students' use of higher-level thinking skills (analysis and inquiry); providing feedback to support students' learning and participation (quality of feedback); and leading discussions that build a deeper understanding of content (instructional dialogue). For example, a coach might suggest that a teacher help deepen students' understanding by applying concepts to the real world. If students appeared to be struggling in a lesson on the metric system, the coach could suggest that the teacher encourage students to discuss items they might buy at the grocery store that come in liters. This discussion could help students gain a practical idea of a liter and where they might find the measurement used in real life.

Less frequently, the coaching addressed dimensions related to classroom management and building supportive relationships with students. On average, teachers in each coaching group spent approximately one cycle focused on CLASS dimensions related to classroom management, such as providing feedback on strategies for managing student behavior (behavior management) and managing instructional time and routines (productivity). Teachers also spent approximately one cycle focusing on dimensions related to building supportive relationships with students. These dimensions included establishing an environment of mutual respect for teachers and students (positive climate); responding to the academic, social, and emotional needs of individual students and the entire class (teacher sensitivity); and incorporating students' interests into classroom activities (regard for student perspective).

Exhibit A.3. Average number of cycles focused on each CLASS dimension and domain, by coaching group

Domain (in bold)/dimension	Average number of cycles as a focus	
	Five-cycle coaching group	Eight-cycle coaching group
Classroom management	**0.9**	**1.3**
Behavior management	0.3	0.6
Productivity	0.7	0.7
Negative climate	0.0	0.0
Building students' understanding	**3.8**	**6.6**
Instructional learning formats	1.2	1.8
Content understanding	1.1	2.2
Analysis and inquiry	1.0	2.1

Domain (in bold)/dimension	Average number of cycles as a focus	
	Five-cycle coaching group	Eight-cycle coaching group
Quality of feedback	1.1	1.9
Instructional dialogue	0.8	1.8
Building supportive relationships with students	**1.0**	**1.0**
Positive climate	0.1	0.1
Teacher sensitivity	0.5	0.5
Regard for student perspective	0.4	0.4
Student engagement	0.1	0.1
Number of teachers	**105**	**102**

Source: Data collected from Teachstone, 2018–2019 school year.

Note: Coaches focused on two different CLASS dimensions in each coaching cycle.

CLASS = Classroom Assessment Scoring System.

Within the CLASS dimensions, coaches focused on specific teaching practices to improve more detailed aspects of teachers' instruction. Consistent with the coaching's primary focus on building students' understanding, the ten most common types of teaching practices addressed by the coaching were in that domain (Exhibit A.4). The two most common practices were facilitation of higher-order thinking and metacognition. Facilitation of higher-order thinking includes providing opportunities for students to engage in activities to identify and investigate problems, examine and interpret data or information, make predictions or hypotheses, and develop arguments or provide explanations. Metacognition includes providing opportunities for students to explain their own thought process, evaluate their own thinking, reflect on and plan their own learning, and model their thought process by thinking out loud.

Exhibit A.4. Ten most frequently covered teaching practices (CLASS indicators) in the coaching cycles, 2018–2019 school year

Teaching practice (CLASS indicator)	CLASS dimension	CLASS domain	Percentage of cycles as a focus	
			Five-cycle coaching group	Eight-cycle coaching group
Facilitation of higher-order thinking (providing opportunities to use higher-order thinking skills)	Analysis and inquiry	Building students' understanding	16	16
Metacognition (modeling the thinking process and helping students explain their thinking)	Analysis and inquiry	Building students' understanding	16	16
Distributed talk (leading classroom discussions that involve many students)	Instructional dialogue	Building students' understanding	13	15
Variety of modalities, strategies, and materials (using multiple approaches to teach lesson content)	Instructional learning formats	Building students' understanding	12	13

Teaching practice (CLASS indicator)	CLASS dimension	CLASS domain	Percentage of cycles as a focus	
			Five-cycle coaching group	Eight-cycle coaching group
Depth of understanding (helping students gain a deeper understanding of content and how facts link to broader concepts or ideas)	Content understanding	Building students' understanding	9	13
Background knowledge and misconceptions (connecting new ideas to students' prior knowledge)	Content understanding	Building students' understanding	12	12
Building on student responses (expanding on and clarifying students' responses or having students expand on or clarify each other's responses)	Quality of feedback	Building students' understanding	10	13
Feedback loops (asking a series of follow-up questions to extend students' thinking or encouraging back-and-forth exchanges among students)	Quality of feedback	Building students' understanding	12	11
Scaffolding (providing support when students struggle with a concept or having other students provide support)	Quality of feedback	Building students' understanding	10	11
Facilitation strategies (asking open-ended questions and actively listening to facilitate productive conversations among students)	Instructional dialogue	Building students' understanding	12	11
Number of cycles			**399**	**671**

Source: Data collected from Teachstone, 2018–2019 school year.

CLASS = Classroom Assessment Scoring System.

A.2.2 Structure of the coaching

The study's coaching consisted of two primary components: (1) an in-person orientation and (2) a set of coaching cycles.

In-person orientation. The in-person orientation occurred before or just after the start of the school year in each study district. Teachers who were unavailable for the in-person orientation attended via webinar. During the orientation, study coaches presented an overview of the teaching practices covered by the coaching and the coaching process and met informally with the participating teachers (Exhibit A.5). The study team also gave an overview of the study and the video recording process. Teachers received copies of a handbook describing the coaching and a guide describing the teaching practices covered by the CLASS.

Exhibit A.5. Amount of time spent on various topics during teacher orientations, 2018–2019 school year

Orientation component	Amount of time spent (minutes)
Overview of teaching practices covered by the coaching	55
Overview of the coaching process	48
Informal coach-teacher interactions	16
Overview of the study and video recording procedures	24
Total	143

Source: Observations of teacher trainings in 14 study districts.

Note: Averages are based on the 14 districtwide in-person orientations conducted for teachers in schools assigned to the coaching groups. Observers documented the minutes spent on each orientation component. Orientations conducted via individual or small-group webinars (for teachers unable to attend the in-person trainings) are not included in these results.

Coaching cycles. The coaching cycles were designed to be collaborative and focused on teachers' strengths to build ongoing and supportive relationships between coaches and teachers. After orientation, each teacher in the coaching groups began participating in the assigned set of coaching cycles.

Schools were randomly assigned to receive either five or eight cycles of coaching during the school year. The study tested an eight-cycle version because prior studies of the MyTeachingPartner program suggested that at least eight cycles had a positive impact on students (Allen et al. 2011, 2015). These studies did not test the effect of providing fewer cycles. This study also tested a five-cycle version of the coaching because the study's expert panel recommended five cycles as more feasible for districts and teachers to implement in a single school year.

Each coaching cycle was intended to take approximately three weeks for teachers assigned to the five-cycle coaching group and approximately two weeks for teachers assigned to the eight-cycle group. The coaching cycles were completed during the study school year, with most (more than 90 percent for both the five- and eight-cycle groups) occurring between October and March. On average, teachers assigned to the five-cycle group completed 4.3 coaching cycles during the school year, and teachers assigned to the eight-cycle group completed 7 coaching cycles (Exhibit A.6).

Exhibit A.6. Number of study-provided coaching cycles completed and average length of coaching conferences

	Five-cycle coaching group	Eight-cycle coaching group
Average number of coaching cycles completed	4.3	7.0
Percentage of teachers completing the following number of coaching cycles…		
0	9	7
1	3	<4
2	3	<4
3	0	0
4	0	0
5	85	4

	Five-cycle coaching group	Eight-cycle coaching group
6	0	<4
7	0	0
8	0	84
Average length of each coaching conference (minutes)	41.1	39.8
Average total time spent in coaching conferences (minutes)	178.5	279.6
Number of teachers	**116**	**110**

Source: Data collected from Teachstone and coaching logs, 2018–2019 school year.

Note: A < or > indicates that the exact percentage has been withheld to protect respondent confidentiality in accordance with National Center for Education Statistics statistical standards, but the percentage is less than or greater than the number following the < or > symbol. Sample includes all teachers randomly assigned to the five- or eight-cycle coaching groups.

Each coaching cycle was organized into five steps:

1. The coach and teacher identified two CLASS dimensions to address in the coaching cycle. The study team video recorded 30 minutes of an English language arts or math lesson in the teacher's classroom. During the video, teachers were expected to implement the practices recommended by the coach for the CLASS dimensions being addressed in the cycle.
2. The coach selected three video clips from the recorded lesson, each lasting about one to two minutes. For each clip, the coach provided written feedback and questions intended to prompt the teacher to reflect on the two CLASS dimensions. The coach's feedback was designed to help teachers become better observers of their own teaching.
3. The teacher viewed the video clips and responded to the coach's questions in writing.
4. After receiving the responses, the coach held a 30- to 45-minute virtual conference with the teacher. During the conference, the coach and teacher discussed the video clips, the coach's feedback, and the teacher's reflections on the practices related to the CLASS dimensions. Then they worked together to decide the next two CLASS dimensions they would address and develop an action plan for improving those dimensions during the next cycle.
5. After the coaching conference, the coach sent the teacher a summary of the conference and a written action plan with (a) video clips of exemplar teachers demonstrating specific practices that aligned with the CLASS dimension that was the focus of the next coaching cycle;[3] (b) summaries of the CLASS dimensions that were the focus of the next cycle; and (c) specific strategies (aligned with the CLASS dimension) for the teacher to capture in the video recorded lesson for the next cycle.

Coaches were expected to follow a standardized approach to implement each of these steps in the coaching cycle. There were three key features of the standardized approach to the coaching:

- ***A specific purpose for each video clip.*** The coach selected three video clips in each cycle: Nice Work, Consider This, and Making the Most. Each clip was intended to achieve a specific goal. The Nice Work clip focused teachers on positive aspects of their teaching. The Consider This clip helped teachers make connections between their practices and student actions and behaviors. Finally, the Making the Most clip showed teachers how interactions with students support learning.
- ***Common structure for providing written feedback.*** Teachstone provided coaches with detailed guidance for structuring their written feedback to teachers. For each video clip, the coach provided written feedback that named and described the CLASS dimension that was the focus of the clip and explained how the video clip exemplified the teaching practice. The written feedback also included questions to help teachers reflect on how they used the teaching practice in the video clip.
- ***Four-step process for each conference***. Coaches followed a standard four-step process for each coaching conference. First, the coach checked in on the teacher's well-being and addressed any questions or concerns about the coaching process. Second, the coach reviewed the video clips and the teacher's responses to written feedback. Third, the coach provided guidance on effective teaching practices that the teacher could use in the classroom. Finally, the coach engaged the teacher in a discussion to select and plan for one to two CLASS dimensions to focus on in the next cycle.

A.3 Coach selection, assignment, and training

This section describes how Teachstone selected coaches to deliver the study's coaching, coaches' assignment to study schools, and the training and ongoing support they received from Teachstone.

A.3.1 Coach selection and characteristics

To select coaches for the study, Teachstone drew from its existing network of MyTeachingPartner coaches and posted a job announcement and description on the Teachstone website. Teachstone shared the job announcement and description via the University of Virginia and other university and professional organizations; emails to its network of current coaches, trainees, and other contacts; and job recruitment sites, including LinkedIn and Indeed.

Teachstone specified three primary qualifications for coaches: (1) a master's degree in education or a related field; (2) at least five years of teaching experience in elementary grades; and (3) experience providing professional development or support to education professionals. Consistent with Teachstone's expectations, 60 percent of coaches who delivered the program held an advanced degree, 73 percent had five years of teaching experience in elementary education, and all of the coaches had experience providing professional development or coaching to teachers (Exhibit A.7).

Exhibit A.7. Characteristics of the study coaches

Characteristic	Percentage[a]
Percentage holding a master's degree in education or a related field	60
Percentage with at least 5 years of teaching experience at elementary school level	73
Percentage with experience providing professional development or support to teachers	100
Percentage with previous experience as a coach	100
Number of coaches	**15**

Source: Authors' compilation based on coach resumes and data collected from Teachstone, 2018–2019 school year.

[a] The sample includes coaches who provided at least two cycles of coaching to participating teachers, including one coach who left the study in October 2018. Two coaches are not included in the table (one who left the study before the coaching began and a second who provided one cycle of coaching to six teachers while another coach was on maternity leave).

A.3.2 Coach assignments to study schools

Within each study district, coaches were randomly assigned to the study schools. Random assignment ensured that particular types of coaches were not systematically assigned to particular types of schools (for example, assigning more experienced coaches to the lowest achieving schools). This allowed the study to attribute any differences in teacher and student outcomes across coaches to differences in coach effectiveness rather than to underlying differences in the types of schools to which the coaches were assigned. The study team assigned a single coach to each school, except in cases where that approach would result in a coach being assigned to too many teachers. The coaching provider sought to assign each coach to approximately 14 teachers based on its understanding of how many teachers a coach could reasonably accommodate from its prior experience with the coaching. If a coach could not be assigned to all of the teachers in a school, the study team randomly selected as many teachers from that school as the coach had capacity for, and then randomly assigned the remaining teachers to another coach. Each coach was assigned a roughly equal number of teachers from the five- and eight-cycle group to ensure that the same coaches delivered the coaching to both groups.

A.3.3 Coach training and ongoing support

Teachstone provided an in-person interactive training for the study coaches in the summer before the study school year. The five-day training included two days of training on how to conduct classroom observations using the CLASS and three days guiding coaches through each aspect of a coaching cycle (Exhibit A.8). The training emphasized opportunities to practice the coaching activities by implementing the steps of a mock coaching cycle. After the in-person training, coaches independently completed a full mock coaching cycle and debrief with a Teachstone coaching specialist. In addition, coaches had to complete a certification test to demonstrate their accuracy in using the CLASS rubric to observe classrooms based on video recorded observations.

After the coaching cycles began, the coaching specialist reviewed between one and three cycles of coaching for each coach to assess fidelity to the model. They reviewed the questions for prompting teacher reflection, the conference summary, and the action plan for the selected cycles. The coaching specialist then provided feedback to help coaches reflect on their coaching and make plans for improvement. After these fidelity reviews, the coaches recorded a teacher conference. The coaching specialist then reviewed the recording to check that the coaches followed the intended four-step process for the conferences.

Exhibit A.8. Amount of coach training time spent on key topics, summer 2018

Training topics	Amount of time spent (minutes)
MyTeachingPartner model	40
MyTeachingPartner goals	121
MyTeachingPartner cycle steps (overview)	47
Video selection	143
Prompt writing	225
Teacher responses	79
Conference	110
Summary email and action plan	105
Organization and teacher engagement	55
All training components (total)	**925**

Source: Observation of MyTeachingPartner coach training held in July 2018.

Note: Coaches were trained during a five-day session held in July 2018. The first two days of the session were used to train the coaches to rate video observations using the CLASS rubric and were not observed. All training conducted during the remaining three days was observed. Observers documented the time spent on each training component.

Coaches received ongoing support from Teachstone throughout the study school year. Each coach attended a 60-minute one-on-one video call with an assigned Teachstone coaching specialist every other week. The calls covered coaching updates, troubleshooting of any challenges, and a discussion of the coach's current coaching cycles. During alternating weeks, all study coaches attended a 90-minute group videoconference facilitated by a Teachstone coaching specialist. These calls focused on a series of topics designed to increase coaching competency, including planning effective reflection questions, supporting effective implementation of cycle work, supporting teachers in problem solving, promoting a growth mindset, and improving quality of feedback. The meetings also included time for coaches to brainstorm and problem solve together, collaborating to generate ideas and solutions. In addition, each coach had to demonstrate their accuracy in using the CLASS observation rubric once during the school year by accurately scoring a sample video.

A.4 Implementation support for the coaching

To increase the chances that the video-based coaching program would run smoothly in the participating districts, the study's technical assistance team communicated frequently with Teachstone and district staff to carefully monitor all program activities. The team:

- Reviewed the credentials of Teachstone coaches to confirm they were as consistent as possible with the qualifications established for the study
- Reviewed materials used in the coach training and teacher orientations for completeness and clarity
- Monitored the coach training, teacher orientations, and coaching cycles to ensure the activities were delivered as intended
- Met regularly with Teachstone staff to review data on completed coaching cycles and teacher engagement in the coaching and to identify strategies to re-engage unresponsive teachers

- Met regularly with district staff to review aggregate data on teachers' participation and gather feedback on the coaching program

A.5 Costs of the coaching

The cost of the five-cycle coaching program as implemented for the study was approximately $228 per student, and the cost of the eight-cycle coaching program was approximately $335 per student. The primary cost driver was the $807 per-cycle cost for Teachstone to provide each cycle of coaching, which included hiring, training, compensating, and supporting the coaches. The cost per cycle for each student was slightly lower for the eight-cycle group compared to the five-cycle group ($46 per cycle for each student for the five-cycle group and $42 per cycle for each student for the eight-cycle group), suggesting some savings from a larger number of coaching cycles. Appendix B provides more details on the study's approach to determining program costs.

APPENDIX B. STUDY DESIGN, DATA COLLECTION, AND ANALYTIC METHODS

This appendix provides more details on the study's design, data collection, and analytic methods.

B.1 Study design

The study team recruited districts and schools to participate in the study and randomly assigned schools to receive five cycles of teacher coaching, eight cycles of teacher coaching, or no study-provided coaching. This section describes how the study team selected the districts and schools for the study and randomly assigned schools.[4]

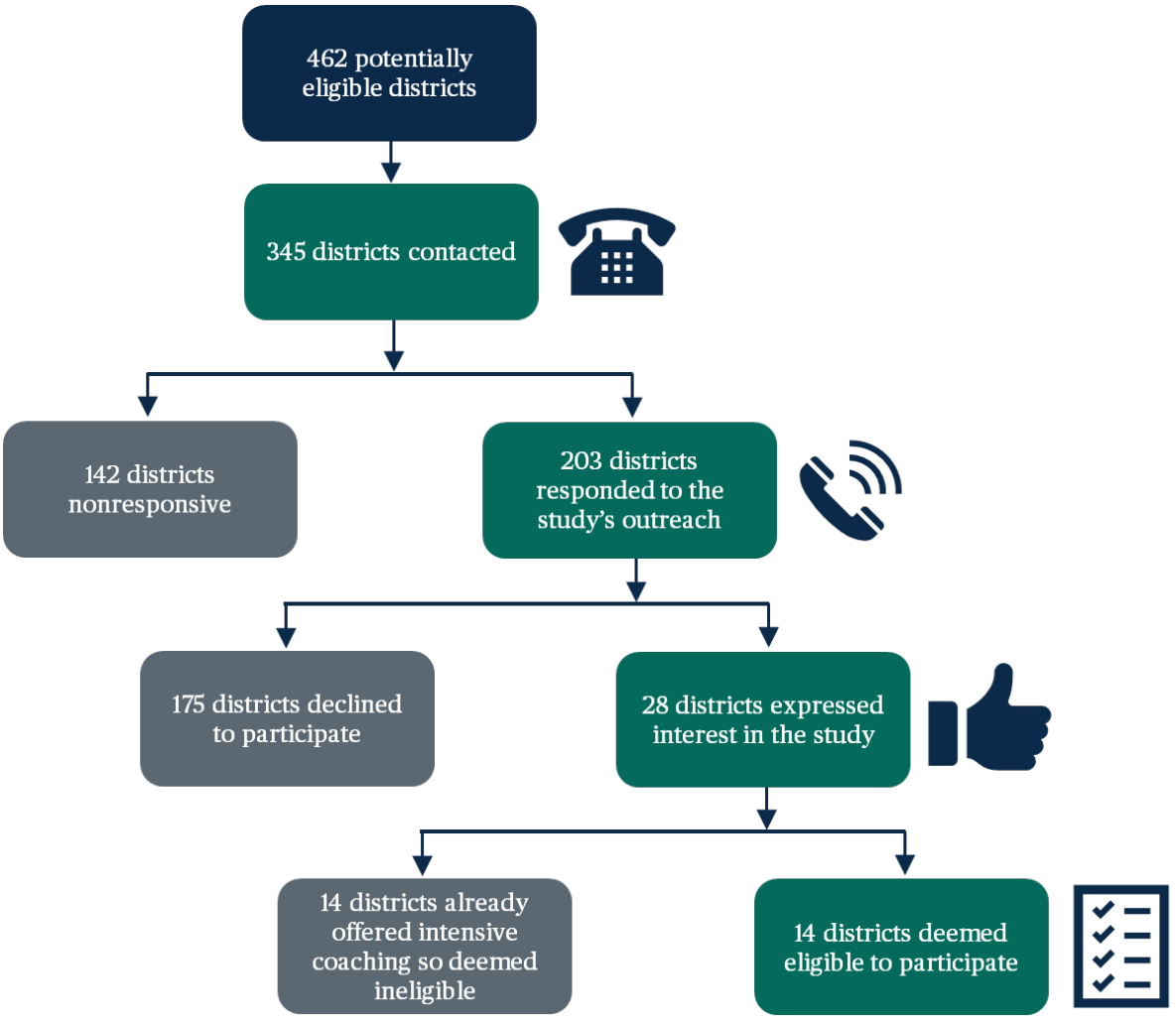
B.1.1 Sample selection and recruitment

The study focused on districts that did not already provide extensive coaching and feedback to their teachers. This ensured that there would be a meaningful contrast between teachers who participated in the study-provided coaching and those who did not. In addition, because the effects of individualized professional development for teachers could differ for elementary and secondary grades, the study focused on teachers of grades 4 and 5.

To efficiently meet the study's sample size requirements, recruitment efforts focused on districts with relatively large numbers of elementary schools. The study team used the U.S. Department of Education's Common Core of Data to identify 462 districts that had at least 17 schools serving grades 4 and 5. To help ensure the sample was geographically diverse, the team classified these districts by U.S. Census Bureau region and prioritized the largest districts in each region for the initial recruitment outreach.

Of the 462 potentially eligible districts, the study team reached out to 345 to assess their interest in and suitability for the study (Exhibit B.1). Ultimately, 28 districts expressed interest in the study. Fourteen of these districts already offered extensive coaching and were therefore excluded from the study. The remaining 14 districts formed the final study sample. Across the 14 participating districts, a total of 107 schools participated in the study.

Exhibit B.1. Results from district recruitment effort



Given the study's focus on districts with at least 17 schools serving grades 4 and 5, study districts differed from typical districts nationwide (Exhibit B.2). For example, study districts were larger, more concentrated in the South, and less concentrated in the Northeast and in the West. They were also more concentrated in suburban and urban areas, more racially diverse, and had smaller shares of students with individualized education programs. Study districts were more similar to large school districts nationwide. However, compared to large school districts nationwide, study districts had smaller shares of English learner students, were more likely to be in the South, and were less likely to be in the Northeast or West.

Study schools also differed from public elementary schools in multiple ways (Exhibit B.3). For example, study schools had higher poverty levels and were larger and more racially diverse than public elementary schools nationwide. Study schools were more similar to schools in larger school districts nationwide (those with at least 17 schools serving grades 4 and 5).

Exhibit B.2. Comparison of study districts and public school districts nationally

Characteristic (percentages[a] unless otherwise noted)	Means			Differences			
				Study districts versus all public school districts		Study districts versus largest public school districts	
	Study districts	All public school districts	Largest public school districts	Difference	*p*-value	Difference	*p*-value
Student racial and ethnic distribution[b]							
Black, non-Hispanic	29	11	18	18*	0.00	11	0.07
Hispanic	21	16	32	5	0.24	-11*	0.01
White, non-Hispanic	41	64	38	-23*	0.00	4	0.55
Other, non-Hispanic	9	8	12	0	0.92	-4*	0.01
Other student characteristics							
Students eligible for free or reduced-price lunch	57	50	54	6	0.30	2	0.73
English language learners	7	7	13	0	0.97	-6*	0.00
Students with Individualized Education Program	12	15	13	-3*	0.00	-1	0.13
District size							
Number of schools (average)	97	6	65	91*	0.03	32	0.43
Number of students (average)	61,800	3,243	42,471	58,556*	0.01	19,329	0.40
District location							
Urban	43	14	52	29*	0.03	-9	0.48
Suburban	50	23	41	27*	0.04	9	0.51
Town	0	16	2	-16*	0.00	-2*	0.00
Rural	7	47	5	-40*	0.00	2	0.73

Characteristic (percentages[a] unless otherwise noted)	Means			Differences			
				Study districts versus all public school districts		Study districts versus largest public school districts	
	Study districts	All public school districts	Largest public school districts	Difference	*p*-value	Difference	*p*-value
Geographic region							
Northeast	0	21	5	-21*	0.00	-5*	0.00
Midwest	7	36	13	-28*	0.00	-6	0.40
South	79	24	46	55*	0.00	32*	0.00
West	14	19	36	-5	0.59	-22*	0.02
Number of districts	**14**	**11,603–15,251**	**469–496**				

Source: Common Core of Data (2017–2018 school year).

Note: Exhibit excludes districts that contain only charter schools. Largest districts are those with at least 17 schools serving grades 4 or 5.

[a] Differences between groups may differ from differences in reported means due to rounding.

[b] Race and ethnicity categories are mutually exclusive.

* Statistically significant at the .05 level, two-tailed test.

Exhibit B.3. Comparison of study schools and public elementary schools nationally

Characteristic (percentages[a] unless otherwise noted)	Mean			Differences			
				Study schools versus all public elementary schools		Study schools versus public elementary schools in largest districts	
	Study schools	All public elementary schools	Public elementary schools in largest districts	Difference	*p*-value	Difference	*p*-value
Student racial and ethnic distribution[b]							
Black, non-Hispanic	36	15	22	21*	0.00	13*	0.00
Hispanic	21	24	34	-4	0.08	-13*	0.00
White, non-Hispanic	36	50	31	-15*	0.00	4	0.17
Other, non-Hispanic	8	10	12	-3*	0.00	-4*	0.00
Students eligible for free or reduced-price lunch	66	56	62	10*	0.00	4	0.16
Number of students (average)	567	463	547	104*	0.00	19	0.29
Student-teacher ratio (average)	17	17	18	0	0.74	0	0.35
Schoolwide Title I status[c]	78	78	73	-1	0.86	5	0.25
Number of schools	**107**	**51,089–54,927**	**19,199–20,248**				

Source: Common Core of Data (2017–2018 school year).

Note: Largest districts are those with at least 17 schools serving grades 4 or 5.

[a] Differences between groups may differ from differences in reported means due to rounding.

[b] Race and ethnicity categories are mutually exclusive.

[c] Schoolwide Title I status refers to schools with student populations that are at least 40 percent low income and that are eligible for Title I funds. This means that the schools are classified by state and federal regulations as high poverty and eligible for additional financial assistance. The 300 largest districts are defined based on number of elementary schools.

* Statistically significant at the .05 level, two-tailed test.

B.1.2 Random assignment

The study team randomly assigned participating schools to one of three groups: a group whose teachers received five cycles of coaching, a group whose teachers received eight cycles of coaching, or a control group that did not receive any study-provided coaching. The primary goal of random assignment was to create groups that were similar at the start of the study in observable and unobservable characteristics. That way, any later differences in outcomes between the three groups could be reliably attributed to the effects of the coaching. Before random assignment, each school selected either their 4th or 5th grade to participate in the study. The study team grouped the schools in each district into sets of three, or "random assignment blocks," based on the similarity of their demographic characteristics (number of students, proficiency rates on state math and English language arts assessments, and share of students who were Black, Hispanic, or eligible for free or reduced-price lunch) and, where possible, whether classes were self-contained or departmentalized and selected grade level for the study. (In 11 out of 29 random assignment blocks, it was not possible to form groups of three schools serving the same grade. As a result, these blocks included one or more schools that had selected grade 4 to participate and one or more schools that had selected grade 5.) The study team then randomly assigned one school in each random assignment block to each group (five coaching cycles, eight coaching cycles, or the control group). This approach helped ensure that the schools in the three groups were similar at the start of the study.

The resulting groups had similar baseline characteristics. Exhibit B.4 shows that students and schools in the three groups had similar student achievement and student demographic characteristics at baseline, although larger shares of students in the control group were eligible for free and reduced-price lunch (72 percent in the control group versus 63 percent in the five-cycle group and 61 percent in the eight-cycle group). In addition, student–teacher ratios were higher among schools assigned to the five-cycle group (18 students per teacher in the five-cycle group, compared with 17 students per teacher in the control and eight-cycle groups). Similarly, Exhibit B.5 shows that teachers in the two coaching groups and the control group had similar years of experience, demographic characteristics, teaching assignments, and teaching practices at the start of the study. Teachers in the five-cycle group were somewhat less likely to have a master's degree than teachers in the other two groups (37 percent compared with 46 percent for teachers in the control group and 43 percent for teachers in the eight-cycle group), although these differences were not statistically significant. Teaching practices at the start of the study were similar across the three groups even among subgroups of teachers defined by years of teaching experience and whether they had weak or strong practices (Exhibits B.6 and B.7).

Exhibit B.4. Comparison of baseline characteristics of students and schools in the control group and five- and eight-cycle coaching groups

Characteristic (percentages unless otherwise noted)	Means				Differences					
	Full sample	Control group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle vs. control		Eight-cycle vs. control		Five-cycle vs. eight-cycle	
					Difference	*p*-value	Difference	*p*-value	Difference	*p*-value
Baseline student achievement (z-scores)										
Math	-0.15	-0.16	-0.14	-0.11	0.02	0.79	0.05	0.41	-0.03	0.60
English language arts	-0.08	-0.13	-0.06	-0.04	0.07	0.32	0.10	0.12	-0.02	0.75
Baseline student characteristics										
Male	50	49	51.00	49.00	2.00	0.12	0.00	0.72	1	0
Race and ethnicity[a]										
Black	38	36	38	36	2	1	-1	1	3	0
Hispanic	22	26	19	23	-7*	0	-4	0	-3	0
White	46	46	46	46	0	1	0	1	0	1
Other	8	6	8	10	2	0	3*	0	-1	0
Eligible for free or reduced-price lunch	67	72	63	61	-9*	0	-11*	0	1	1
English language learner	10	12	9	10	-3	0	-2	0	-1	1
Individualized Education Plan	10	10	10	10	0	1	0	1	0	1
School characteristics										
Number of students (average)	566.78	579.70	573.45	554.81	-6.25	0.84	-24.90	0.42	18.64	0.52
Student-teacher ratio (average)	17.19	16.89	17.87	16.99	0.98*	0.00	0.10	0.76	0.89*	0.00
Schoolwide Title I status[b]	78	73	81	83	8	0	10	0	-2	1
Number of students		**1,492–3,290**	**1,211–2,915**	**1,155–2,701**						
Number of schools		**19–37**	**17–36**	**16–34**						

Source: Student outcomes and characteristics come from student administrative records (2017–2018 school year). School characteristics come from Common Core of Data (2017–2018 school year).

Note: Test scores were converted to z-scores by subtracting the mean and dividing by the standard deviation of scores for all students in that state and grade level. Sample sizes vary due to the availability of baseline data.

[a] Race and ethnicity categories are not mutually exclusive unless the district reported mutually exclusive categories, so percentages may sum to more than 100.

[b] Schoolwide Title I status refers to schools with student populations that are at least 40 percent low income and that are eligible for Title I funds. This means that the schools are classified by state and federal regulations as high poverty and eligible for additional financial assistance.

* Statistically significant at the .05 level, two-tailed test.

Exhibit B.5. Comparison of baseline characteristics of study teachers in the control group and five- and eight-cycle coaching groups

Characteristic (percentages unless otherwise noted[a])	Means				Differences					
					Five-cycle vs. control		Eight-cycle vs. control		Five-cycle vs. eight-cycle	
	Full sample	Control group	Five-cycle coaching group	Eight-cycle coaching group	Difference	*p*-value	Difference	*p*-value	Difference	*p*-value
Years of teaching experience[b]	11.44	11.45	10.96	12.29	-0.49	0.63	0.84	0.42	-1.33	0.26
Race and ethnicity[c]										
Black	28	29	30	27	1	0.79	-1	0.70	3	0.50
Hispanic	6	4	7	9	2	0.22	5	0.13	-2	0.47
White	71	74	67	72	-7	0.13	-2	0.68	-5	0.24
Other	3	3	5	4	2	0.26	0	0.73	1	0.49
Highest degree										
Bachelor's	48	45	57	46	11	0.10	0	0.91	11	0.14
Master's or higher degree	51	55	43	54	-11	0.10	0	0.91	-11	0.14
Grades taught[c]										
4	61	62	62	61	0	0.96	-1	0.82	1	0.78
5	41	40	42	43	2	0.73	3	0.66	-1	0.88
Content areas taught										
Math	74	75	76	75	0	0.83	0	0.92	1	0.73
English language arts	74	80	72	72	-7	0.08	-7	0.07	0	1.00
CLASS rating from classroom observations at the start of the study school year	4.57	4.60	4.55	4.53	-0.06	0.32	-0.08	0.16	0.02	0.73
Teaches a self-contained class	46	49	48	45	-1	0.87	-3	0.59	2	0.73
Number of teachers		**129–132**	**109–112**	**103–107**						

Characteristic (percentages unless otherwise noted[a])	Means				Differences					
					Five-cycle vs. control		Eight-cycle vs. control		Five-cycle vs. eight-cycle	
	Full sample	Control group	Five-cycle coaching group	Eight-cycle coaching group	Difference	*p*-value	Difference	*p*-value	Difference	*p*-value
Number of schools		37	35–36	34						

Source: Baseline teacher participation form in fall 2018; teacher survey administered in spring 2019; CLASS ratings from fall 2018 observations.

Note: Sample sizes vary due to the availability of baseline data.

[a] Differences between groups may differ from differences in reported means due to rounding.

[b] Years of experience include all years of teaching before and including the 2018–2019 school year.

[c] Categories are not mutually exclusive, so percentages may sum to more than 100.

CLASS = Classroom Assessment Scoring System.

Exhibit B.6. Baseline teacher practice ratings in the control group and five- and eight-cycle coaching groups, by teacher experience

CLASS rating from classroom observations at the start of the study school year	Novice teachers			Experienced teachers		
	Control	Five-cycle coaching group	Eight-cycle coaching group	Control	Five-cycle coaching group	Eight-cycle coaching group
	Means	Means	Means	Means	Means	Means
Overall CLASS score	4.49	4.59	4.58	4.64	4.54	4.51
Classroom management	6.34	6.42	6.50†	6.52	6.44	6.33*†
Building students' engagement	5.29	5.38	5.52	5.34	5.35	5.20
Building students' understanding	3.54	3.58	3.47	3.58	3.43	3.49
Building supportive relationships with students	3.94	4.16	4.19	4.29	4.21	4.15
Number of teachers	**34**	**32**	**32**	**91**	**69**	**69**
Number of schools	**19**	**21**	**21**	**35**	**29**	**31**

Source: CLASS ratings from fall 2018 observations.

Notes: Novice teachers are those who have been teaching for five years or less; experienced teachers are those who have been teaching for more than five years. Differences in baseline scores between teachers in the five-cycle and eight-cycle groups, and between novice and experienced teachers are not statistically significant at the .05 level, two-tailed test.

* Statistically significant difference between the coaching group teachers and the control group teachers at the .05 level, two-tailed test.

† Statistically significant difference in impacts between novice and experienced teachers at the .05 level, two-tailed test.

CLASS = Classroom Assessment Scoring System.

Exhibit B.7. Baseline teacher practice ratings in the control group and five- and eight-cycle coaching groups, by quality of teachers' practices at baseline

CLASS rating from classroom observations at the start of the study school year	Teachers with weaker teaching practices at baseline			Teachers with stronger teaching practices at baseline		
	Control	Five-cycle coaching group	Eight-cycle coaching group	Control	Five-cycle coaching group	Eight-cycle coaching group
Outcome	**Means**	**Means**	**Means**	**Means**	**Means**	**Means**
Overall CLASS score	3.99	4.05	3.96	5.15	5.05*	5.07
Classroom management	6.11	6.08	6.10	6.75	6.70	6.66
Building students' engagement	4.69	4.87	4.75	5.82	5.77	5.69
Building students' understanding	2.94	2.96	2.85	4.16	4.02	4.12
Building supportive relationships with students	3.38	3.56*†	3.42	4.98	4.89†	4.86
Number of teachers	**38**	**38**	**34**	**44**	**34**	**31**
Number of schools	**24**	**23**	**24**	**26**	**22**	**23**

Source: CLASS ratings from fall 2018 observations.

Notes: Quality of teachers' teaching practices is defined based on teachers' baseline CLASS scores. The CLASS ranges from 1 to 7. Teachers with weaker teaching practices at the start of the study are those who scored in the bottom third of CLASS scores for the sample, and those with stronger teaching practices at the start of the study are those who scored in the top third. Differences in baseline scores between teachers in the five-cycle and eight-cycle groups, and between teachers with weaker and stronger practices at baseline are not statistically significant at the .05 level, two-tailed test.

* Statistically significant difference between the coaching group teachers and the control group teachers at the .05 level, two-tailed test.

† Statistically significant difference in impacts between novice and experienced teachers at the .05 level, two-tailed test.

CLASS = Classroom Assessment Scoring System.

B.2 Data collection

To assess the effects of the study-provided coaching and describe how it was implemented, the study team collected data from several sources. Exhibit B.8 lists these data sources. Exhibit B.9 lists the response rates for the data sources used to measure the effects of the coaching.

Exhibit B.8. Data sources

Data source	Data obtained	Timing of data collected	Respondent
Data to measure effects (collected for coaching and control groups)			
Student records[a]	Student achievement and background characteristics from the baseline (2017–2018) and study (2018–2019) school years	Fall 2018 and fall 2019	Students
Teacher participation form	Years of experience, grade(s) and subject(s) taught, feelings of preparedness for teaching	Summer 2018	Teachers
Teacher survey[b]	Teacher perceptions of the amount, quality, and usefulness of feedback received	Spring 2019	Teachers
Classroom observations[c]	Quality of teachers' general classroom practices, as measured by the Classroom Assessment Scoring System (CLASS) rubric	Fall 2018 and spring 2019	Not applicable (study team observed teachers)
	Quality of teachers' English language arts-specific classroom practices, as measured by the Protocol for Language Arts Teaching Observations (PLATO)[d]		
Data to measure implementation (collected for coaching groups)			
Coach resumes	Coach's prior coaching and teaching experience	Summer 2018	Coaches
Online coaching platform	Number of coaching cycles partially and fully completed, length of each coaching cycle, and teaching practices covered in each cycle	Fall 2018–spring 2019	Coaches
Coach feedback logs	The length, content, and structure of each coaching conference	Fall 2018–spring 2019	Coaches
Teacher orientation observations	Time spent on various topics during orientation	Fall 2018–spring 2019	Not applicable (study team observed teacher orientations)
Coach training observations	Time spent on various topics during coach training	Fall 2018–spring 2019	Not applicable (study team observed coach training)
Recordings of coaching conferences	The content and quality of a random sample of six coaching conferences for each coach. Sampled conferences were recorded and rated using the Coach Quality Checklist—a rubric consisting of 26 items. Ratings for each conference were then averaged across coaches to determine the average quality of each coach's conferences.[e]	Fall 2018–spring 2019	Not applicable (study team coded the conference recordings)

Data source	Data obtained	Timing of data collected	Respondent
Data to describe the study sample (collected for study districts and schools and all districts and schools nationwide)			
Common Core of Data	Characteristics of study districts and schools	2017–2018	Districts and schools
ED*Facts* Achievement Results for State Assessments in Mathematics and Reading/Language Arts	The percentage of students in study schools scoring at or above the state-defined proficiency level on the state math and reading/language arts assessments	2016–2017	States

[a] Most teachers (97 percent in the five-cycle group and 88 percent in the eight-cycle group) completed their last coaching cycle before students took the state assessment on which the student achievement data are based.

[b] Data from the teacher survey were used to describe the implementation and effects of the study-provided coaching on teacher outcomes. The teacher survey was administered during the same time frame to all teachers and did not vary for teachers in the two coaching groups or the control group.

[c] Almost all teachers (100 percent in the five-cycle group and 96 percent in the eight-cycle group) completed their last coaching cycle before the study team conducted their spring classroom observations.

[d] PLATO scores were not included in the original study design. They were added to further explore the coaching's effects on teachers' English language arts–specific practices after the study found that five cycles of coaching improved students' English language arts achievement.

[e] If one of the randomly selected conferences was the last conference with a teacher, it was omitted from this average, because some items on the Coach Quality Checklist are not applicable in the final session.

Exhibit B.9. Response rates for data sources used to estimate effects of the study-provided coaching

Data collected	Overall	Control group	Five-cycle coaching group	Eight-cycle coaching group
Student records				
Student-level responses				
Math scores	94.3	94.5	94.9	93.4
English language arts scores	95.0	95.3	95.2	94.5
School-level response rates				
Math scores	100.0	100.0	100.0	100.0
English language arts scores	100.0	100.0	100.0	100.0
Teacher records				
Teacher-level responses				
Classroom observations	91.2	94.7	88.4	89.9
Teacher survey	98.0	99.2	99.1	95.4
School-level responses				
Classroom observations	93.5	97.3	91.7	91.2
Teacher survey	100.0	100.0	100.0	100.0

Source: Student records, teacher survey, and classroom observations from spring 2019.

B.3 Analytic methods

This section describes the study team's approach to examining the effects of study-provided coaching on student achievement and teachers' practices. It first describes the construction of outcome measures for the study. Next, it provides details about the study's analytic methods, including the methods used to estimate the effects of the coaching on these outcomes and the methods used to estimate the relationship between the characteristics of the coaching and its effects on teachers' practices and student achievement. Finally, it discusses estimation of the cost effectiveness of the coaching.

B.3.1 Constructing outcome measures

This section discusses the methods used to construct measures of teachers' practices and student achievement.

Measures of student achievement. To measure student achievement, the study team used students' test scores on state assessments in math and English language arts, standardized across the different states in the study. To standardize, the test scores were converted to z-scores by subtracting the statewide mean and dividing by the statewide standard deviation for that year, grade, and subject. After estimating effects on the standardized scores, the estimates were converted into test score percentiles to make them easier to interpret. To calculate these percentiles, the team determined the mean student achievement in z-score units for students taught by teachers in the control group and for students taught by teachers in the coaching groups. The team then calculated the percentiles based on the proportion of the area under the normal curve below these z-score

values. The team converted the effects on student achievement into average months of learning by dividing the effects by the average one-year gain in achievement on nationally normed assessments for grades 4 and 5.[5]

Measures of teachers' practices. The study team measured teachers' general practices using the Classroom Assessment Scoring System (CLASS) rubric. The CLASS rubric is an established measure of general teaching practices that has evidence of reliability and validity.[6] After the study found that five cycles of coaching improved students' English language arts scores, the study team also coded the videos with a rubric that measures teachers' English language arts–specific practices using the Protocol for Language Arts Teaching Observations (PLATO) rubric to examine whether the study's coaching affected these practices. The PLATO also has evidence of reliability and validity.[7] The CLASS includes three broad domains of practices that are made up of finer-grained practices referred to as "dimensions" (Exhibit B.10). It also includes one dimension (building student engagement) that is measured separately and not part of the three domains. The PLATO includes four broad domains of practices made up of finer-grained practices referred to as "elements." To construct domain-level scores of the CLASS, the study team calculated a simple average of the dimension-level scores associated with each domain. Similarly, to construct domain-level scores of the PLATO, the study team calculated a simple average of the element-level scores associated with each domain. To construct overall scores for both rubrics, the study team calculated a simple average of the dimension- or element-level scores for each.

Although both the CLASS and PLATO have evidence of reliability from prior studies, the study team also calculated the reliability based on the scores used in this study to ensure the rubric scores provided a consistent measure of teachers' practices across teachers and raters. The team measured the extent to which the dimensions that make up each domain measure a common aspect of teaching (internal consistency reliability) using two measures: Chronbach's alpha and McDonald's omega.[8] Both of these measures range from 0 to 1 and describe the extent to which the overall domain score is correlated with the dimension scores that make up that domain. Chronbach's alpha assumes that every dimension measures the overall domain score with the same level of precision, while McDonald's omega is more flexible because it allows each dimension to measure the overall domain with different levels of precision. Reliabilities above 0.7 are generally considered acceptable.[9]

Exhibit B.11 shows that the overall CLASS score and domain scores have high internal consistency reliability, with reliabilities above 0.75 for the baseline videos of teachers' classrooms taken in the fall and the follow-up videos taken in the spring. The overall PLATO score and the disciplinary demand domain have acceptable reliability (above 0.7), but the other domains have low reliabilities (between 0.45 and 0.67) (Exhibit B.12).

Exhibit B.10. Domains and associated dimensions/elements of the CLASS and PLATO rubrics

CLASS domains and associated dimensions
Classroom management
Behavior management
Productivity
Negative climate
Building students' understanding
Instructional learning formats
Content understanding
Analysis and inquiry
Quality of feedback
Instructional dialogue
Building supportive relationships with students
Positive climate
Teacher sensitivity
Regard for student perspectives
Building student engagement (stand-alone dimension)
PLATO domains and associated elements
Instructional scaffolding
Modeling and use of models
Strategy use and instruction
Feedback
Accommodations for language learning
Disciplinary demand
Intellectual challenge
Classroom discourse
Text-based instruction
Classroom environment
Behavior management
Time management
Representations and use of content
Representation of content
Connections to prior academic knowledge
Purpose

Note: Domains are in bold with associated dimensions (for the CLASS) or elements (for the PLATO) below each domain.

CLASS = Classroom Assessment Scoring System; PLATO = Protocol for Language Arts Teaching Observations.

Exhibit B.11. Internal consistency reliability of the CLASS rubric

Domain	Dimensions	Chronbach's alpha		McDonald's omega	
		Baseline videos	Follow-up videos	Baseline videos	Follow-up videos
Classroom organization	Behavior management, negative climate, productivity	0.79	0.78	0.84	0.81
Emotional support	Positive climate, regard for student perspectives, teacher sensitivity	0.80	0.81	0.82	0.82
Instructional support	Instructional dialogue, quality of feedback, instructional learning formats, analysis and inquiry, content understanding	0.86	0.87	0.86	0.87
Student engagement	Student engagement	n.a.	n.a.	n.a.	n.a.
Overall CLASS score		**0.85**	**0.84**	**0.91**	**0.84**

Source: Classroom observations from fall 2018 and spring 2019.

Note: The student engagement domain includes a single item so internal consistency reliability cannot be computed for this domain.

CLASS = Classroom Assessment Scoring System, n.a. = not applicable.

Exhibit B.12. Internal consistency reliability of the PLATO rubric

Domain	Dimensions	Cronbach's alpha		McDonald's omega	
		Baseline videos	Follow-up videos	Baseline videos	Follow-up videos
Instructional scaffolding	Modeling and use of models, strategy use and instruction, feedback, accommodations for language learning	0.66	0.63	0.67	0.64
Disciplinary demand	Intellectual challenge, classroom discourse, text-based instruction	0.64	0.74	0.65	0.76
Classroom environment	Behavior management, time management	0.50	0.45	0.63	0.52
Representations and use of content	Representation of content, connections to prior academic knowledge, purpose	0.60	0.59	0.66	0.62
Overall PLATO score		**0.71**	**0.71**	**0.73**	**0.73**

Source: Classroom observations from fall 2018 and spring 2019.

Note: The student engagement domain includes a single item so internal consistency reliability cannot be computed for this domain.

PLATO = Protocol for Language Arts Teaching Observations.

The study team also measured the extent to which different coders used the rubric similarly when observing the same lesson (inter-rater reliability) to ensure the scores provided a consistent measure of teachers' practices. After coding a set of 10 videos, each coder scored a calibration video to identify and address any drift in their scores over time. Because all coders and a master coder (an expert in the use of the rubric) coded the same calibration videos, the study could compare coders' scores with each other and with the master coder. To assess inter-rater reliability for the CLASS and PLATO rubrics, the study examined three measures based on these calibration videos: the percent agreement across dimensions, the linearly weighted Cohen's kappa, and the linearly weighted Gwet's AC1.

The first measure, the percentage of dimensions where coders assigned the same or adjacent scores to the same classroom observation video, aligns with the study's approach to measuring reliability of coders during the coding process with calibration videos. The developer of the CLASS required that coders assign the same score or an adjacent score as the master coder on 80 percent of the dimensions. CLASS coders were not permitted to continue coding if they failed to meet this threshold for three calibration videos in a row. Similarly, the study team required PLATO coders to achieve exact agreement with the master coder on half of the dimensions and exact or adjacent agreement on 90 percent of the dimensions. PLATO coders were not permitted to continue coding if they failed to meet this threshold for 5 out of 6 consecutive calibration videos.

The other two measures, Cohen's kappa and Gwet's AC1, are on a scale of 0 to 1 and measure the extent to which coders assign the same dimension scores but accounts for the possibility that raters assign the same score by chance. Although Cohen's kappa is commonly used, it may be biased, and Gwet's AC1 can avoid this potential bias.[10] The weighting of both statistics takes into account the degree of any disagreement between coders–giving greater weight to dimensions with smaller discrepancies in coder-assigned scores when calculating the reliability (Gwet 2012). Although there are not exact cutoffs for defining low, moderate, and high inter-rater reliability for

classroom observation rubrics, reliabilities above 0.8 are generally considered high, reliabilities between 0.6 to 0.8 are considered moderate, and reliabilities below 0.6 are considered low.[11]

For all three measures, the exhibits show inter-rater reliability in terms of coders' agreement with the master coder and coders' agreement with each other. Coders' level of agreement with the master coder shows how accurately coders were using the rubric to score videos, while coders' level of agreement with each other shows how consistently coders used the rubric.

The rubric scores had high inter-rater reliability based on the measure that aligned with the expectations of the rubric developers—the percent exact or adjacent agreement. Coders assigned the same score or an adjacent score as the master coder 88 percent of the time on the CLASS and 94 percent of the time on the PLATO.

The coders had lower inter-rater reliability based on Cohen's kappa and Gwet's AC1, however the levels were comparable to those from other studies that use classroom observation rubrics. The CLASS coders had moderate levels of inter-rater reliability based on Cohen's kappa and Gwet's AC1 (mean values range from 0.57 to 0.72), while the PLATO coders had low to moderate levels of inter-rater reliability based on these measures (mean values range from 0.39 to 0.65). These values are similar to, or better than, inter-rater reliability for other studies. For example, the CLASS technical manual documents kappas ranging from 0.07 to 0.31 across three studies.[12] A review of published and unpublished studies found an average Cohen's kappa of 0.54 across six studies that used classroom observation rubrics, with values ranging from 0.34 to 0.72.[13]

Exhibit B.13. Inter-rater reliability of the CLASS rubric

Method	Comparison	Mean	Median
Linearly weighted Cohen's kappa	Master coder and each coder within each video	0.60	0.61
	All coders except master within each video	0.57	0.56
Linearly weighted Gwet's AC1	Master coder and each coder within each video	0.72	0.72
	All coders except master within each video	0.72	0.72
Percent exact or adjacent agreement	Master coder and each coder within each video	0.88	0.89
	All coders except master within each video	0.87	0.86

Source: Classroom observations from fall 2018 and spring 2019.

Note: The reliability calculations are based on the calibration videos that coders completed after coding every 10 videos.

CLASS = Classroom Assessment Scoring System.

Exhibit B.14. Inter-rater reliability of the PLATO rubric

Method	Comparison	Mean	Median
Linearly weighted Cohen's Kappa	Master coder and each coder within each video	0.50	0.47
	All coders except master within each video	0.39	0.42
Linearly weighted Gwet's AC1	Master coder and each coder within each video	0.65	0.67
	All coders except master within each video	0.58	0.60
Percent exact or adjacent agreement	Master coder and each coder within each video	0.94	0.96
	All coders except master within each video	0.92	0.92

Source: Classroom observations from fall 2018 and spring 2019.

Note: The reliability calculations are based on the calibration videos that coders completed after coding every 10 videos.

PLATO = Protocol for Language Arts Teaching Observations.

The study team also examined the amount of variation in CLASS scores that was due to variation across teachers, rather than other factors, such as the subject area of the lesson, the type of lesson video recorded, and the coder assigned to score the video. The team calculated the intra-class correlations (ICCs) separately for baseline videos and follow-up videos, using a hierarchical linear model to distinguish the variance in CLASS and PLATO scores due to teachers, videos, and segments. Because each video was coded by a different rater and captured a different lesson, the variance due to videos also captures variation due to raters and lessons.

The ICCs show that a relatively small amount of variation in the overall CLASS scores is due to differences across teachers—21 percent for the baseline videos and 15 percent for the follow-up videos (Exhibit B.15). A variety of other factors—such as the type of lesson that was video recorded, the subject area being taught in the video, and the rater assigned to code the video—account for 65 percent of the variation in baseline videos and 75 percent of the variation in follow-up videos. Some of this variation due to other factors may reflect true variation across teachers (for example, a teacher may have stronger teaching practices when teaching math compared to English language arts).

These ICCs are similar to those found in other studies that analyzed CLASS scores. For example, an IES study of performance feedback for teachers found that differences across teachers accounted for 24 percent of the variation in overall CLASS scores in the study's first year and 33 percent of variation in the second year (Garet et al. 2017). The large-scale Measures of Effective Teaching study found that differences across teachers accounted for 31 percent of variation in overall CLASS scores (Kane and Staiger 2012).

The ICCs for the PLATO scores are lower than those for the CLASS. The amount of variation in PLATO scores due to differences across teachers is 4 percent for the baseline videos and 9 percent for the follow-up videos. The low ICCs for the PLATO may limit the study's ability to measure the impacts of the coaching on teachers' PLATO scores.

Exhibit B.15. Intra-class correlations for overall CLASS scores

Video type	Intra-class correlation
Baseline videos	0.21
Follow-up videos	0.15

Source: Classroom observations from fall 2018 and spring 2019.

CLASS = Classroom Assessment Scoring System.

Exhibit B.16. Intra-class correlations for overall PLATO scores

Video type	Intra-class correlation
Baseline videos	0.04
Follow-up videos	0.09

Source: Classroom observations from fall 2018 and spring 2019.

PLATO = Protocol for Language Arts Teaching Observations.

B.3.2 Estimating effects

To estimate the effects of the coaching on student achievement and teachers' practices, the study team used the following model:

$$y_{ijb} = \alpha + \delta_8 T_{jb8} + \delta_5 T_{jb5} + \beta X_{jb} + \gamma Z_{ijb} + \phi B_b + e_{ijb}$$

where y_{ijb} is the outcome of interest for student or teacher i in school j and random assignment block b; α is an intercept term; T_{jb8} is an indicator equal to one if school j in block b was assigned to the group that received eight cycles of coaching, and zero otherwise; T_{jb5} is an indicator equal to one if school j in block b was assigned to the group that received five cycles of coaching, and zero otherwise; X_{jb} is a vector of school-level covariates measured at the start of the school year; Z_{ijb} are individual-level covariates measured in the year prior to the study school year; and B_b is a vector of indicators for random assignment blocks. The parameter ϕ is a vector of random assignment block fixed effects; and e_{ijb} is an individual-level error term. The parameter δ_8 captures the average effect on the outcome for teachers or students of teachers assigned to eight cycles of coaching, relative to the business-as-usual control group. The parameter δ_5 captures the average effect on the outcome for teachers or students of teachers assigned to five cycles of coaching.[14]

The study team estimated Equation 1 using ordinary least squares and used Huber-White sandwich standard errors to account for clustering at the school level. The team also calculated the unadjusted mean outcomes for the control group and mean outcomes for both coaching groups (the unadjusted mean outcomes for the control group plus the average effect of either the five- or eight-cycle coaching groups).

Covariates. All models controlled for random assignment block fixed effects to reflect the blocked random assignment design. The models also controlled for additional covariates to improve the precision of estimates and to account for any chance imbalances between the groups. The models to estimate effects on both student achievement and teachers' practices included controls for school- and teacher-level covariates. The models for

student achievement also included controls for student-level covariates. Exhibit B.17 summarizes the covariates included in these models.

Exhibit B.17. Baseline covariates included in models used to estimate effects of the coaching

Baseline covariates	Outcome variables	
	Student achievement	Teachers' classroom practices
School-level covariates		
School enrollment	X	X
Teacher-student ratios	X	X
School-level averages of student level covariates	X	X
Teacher level covariates		
Years of teaching experience	X	X
CLASS scores	X	X
Feelings of preparedness for teaching	X	X
Student level covariates		
Math and English language arts scores	X	
Gender	X	
Race and ethnicity	X	
Free or reduced-price lunch eligibility	X	
English learner status	X	
Individualized Education Plan	X	
Random assignment level covariates		
Random assignment block fixed effects	X	X

Weights. The study weighted student and teacher outcomes so that each school contributed equally to the average effect estimate. That is, the study assigned weights to individuals with non-missing outcomes so that the sum of their weights was equal across all schools. An individual i in school j was weighted by $W_{ij} = 1/N_j$, where N_j was the number of individuals with non-missing values of the outcome for school j. With these weights, each school received the same weight in the analysis, regardless of the number of students or teachers in the school. This ensured the results were not overly influenced by the effects of the coaching in larger schools.

Treatment of missing data. The analysis included only individuals who had non-missing values of the outcome variables; individuals with missing values of an outcome variable were excluded from the estimation of effects on that outcome. Individuals were not excluded from the analysis samples if they had missing covariate values, as long as they had non-missing values of the outcome. For each covariate used to estimate the effects of the coaching, the study team replaced missing values with a placeholder value (zero) and included a binary indicator for whether the observation had a missing value. Simulations have shown that this approach to handling missing covariate data is likely to keep estimation bias at less than 0.05 standard deviations.[15]

Samples. For analyses of teacher outcomes, the study team defined the sample to be all eligible teachers who were in study schools and grades at the time of data collection. All teachers in participating grades and schools were eligible unless they were only teaching English learners, they were only teaching special education students, or their district asked to exclude them from the study prior to random assignment. One district chose to exclude teachers with one or two years of experience prior to random assignment; another excluded particular teachers (because, for example, they were already receiving coaching or were teaching gifted and talented students). Teachers who opted not to participate after random assignment were considered eligible and were included in the models to estimate effects. For analyses of student outcomes, the study team defined the sample as students who were enrolled in a study school and taught by an eligible teacher at the beginning of the year.

Estimation of effects for subgroups. The study team estimated the effects of the program on different subgroups of teachers based on their years of teaching experience and teaching practice scores at the start of the study (Exhibit B.18). These subgroup models estimate the causal effect of being assigned to receive five or eight cycles of coaching among teachers in the respective subgroups and their students.

To estimate effects for subgroups defined by teachers' experience and baseline practice scores, the study team estimated a modified version of Equation 1 that adds an indicator for being in the subgroup and an interaction between that indicator and both coaching group indicators. That is, the team estimated the following model:

$$y_{ijb} = \alpha + \delta_8 T_{jb8} + \delta_5 T_{jb5} + \pi_1 Group2_{ijb} + \pi_2 \left(T_{jb8} \times Group2_{ijb} \right) + \pi_3 \left(T_{jb5} \times Group2_{ijb} \right) + \beta X_{jb} + \gamma Z_{ijb} + \phi_b + e_{ijb}$$

where $Group2_{ijb}$ represents one of the two subgroups, and $Group1_{ijb}$ is the omitted category. In this model, the effects of eight cycles of coaching on subgroup 1 and 2 are δ_8 and $\delta_8 + \pi_2$, respectively.

Exhibit B.18. Subgroups examined

Subgroups	Definition
Teacher experience level	
Novice	The teacher had five or fewer years of teaching experience
Experienced	The teacher had more than five years of teaching experience
Baseline teacher practices	
Teachers with weaker practices at baseline	The teacher had a baseline CLASS score in the bottom third of the sample
Teachers with stronger practices at baseline	The teacher had a baseline CLASS score in the top third of the sample

B.3.3 Estimating the relationship between the coaching's effects on teachers' practices and its effects on student achievement

The study team conducted a correlational analysis to better understand the relationships between the coaching's effects on teachers' practices and its effects on student achievement. This analysis included two steps.

In the first step, the study team estimated effects on student achievement and teachers' practices at the random assignment block-level using the following modified version of Equation 1:

$$y_{ijb} = \sum_{b=1}^{B}\left(\delta_{b8}T_{jb8}B_j^b + \delta_{b5}T_{jb5}B_j^B\right) + \phi_b + e_{ijb}$$

where B_j^b is an indicator equal to 1 if school j is in block b and zero otherwise; δ_{b8} is the effect of being assigned to receive eight cycles of coaching in block b; δ_{b5} is the effect of being assigned to receive five cycles of coaching in block b; and all other terms are as defined above. This model excludes covariates in order to conserve degrees of freedom.

In the second step, the study team estimated a series of bivariate correlations between the coaching's block-specific effects on teachers' practices and student achievement. On average, each block included two coaches.

B.3.4 Estimating the cost effectiveness of the coaching

To determine the cost effectiveness of the coaching program, the study team first identified the key components, or ingredients, of implementation and how much of each ingredient was needed to provide five or eight cycles of coaching. The team then used national price data from the Center for Benefit-Cost Studies of Education (CBCSE) to determine the total cost of each ingredient.[16] For ingredients not included in CBCSE data, the study team obtained cost information directly from Teachstone. This information included Teachstone's costs for the teacher orientation, each coaching cycle, and the Teachstone camera kit teachers would typically use to record their instruction for the coach. (For the study, study staff recorded instruction using camera kits provided by the study. However, to better capture the costs that districts would typically face in implementing this type of program, the study team instead considered the cost of the Teachstone camera kits for this analysis.)

The cost of the five-cycle coaching program as implemented for the study was approximately $228 per student, and the cost of the eight-cycle coaching program was approximately $335 (Exhibit B.19). The primary cost driver was the $807 per-cycle cost for Teachstone to provide each cycle of coaching, which included hiring, training, compensating, and supporting the coaches.

The study compared the cost effectiveness of five cycles of coaching to other education strategies with rigorous studies showing their effectiveness for improving student achievement. (It did not compare the costs of eight cycles of coaching because it did not find evidence that eight cycles improved student achievement.) The comparisons included three strategies: teacher pay-for-performance, class size reduction, and transfer incentives for high-performing teachers. The study focused on these strategies for comparison because, like the study's coaching, they all (1) seek to improve student achievement by influencing teachers' effectiveness, (2) could plausibly be implemented in grades 4 and 5, and (3) have rigorous evidence of effectiveness and detailed information on costs from existing studies. As shown in Exhibit B.20, five cycles of coaching has a lower cost per

unit increase in student achievement compared to teacher pay-for-performance, class size reduction, and incentives for high-performing teachers to transfer to low-performing schools.[17]

Exhibit B.19. Costs of providing five and eight cycles of coaching, by ingredient

Ingredient	Quantity (five cycles)	Quantity (eight cycles)	Price per unit	Cost (five cycles)	Cost (eight cycles)
Personnel					
Teachers' time in orientation	4 hours for each of 115 teachers	4 hours for each of 107 teachers	$66[a] per hour	$30,535	$28,411
Teachers' time in coaching cycles	1 hour per cycle for 5 cycles for each of 115 teachers	1 hour per cycle for 8 cycles for each of 107 teachers	$66[a] per hour	$38,169	$56,821
Teachstone orientation costs	One orientation for each of 14 districts	One orientation for each of 14 districts	$4,044[b] per orientation	$56,612	$56,612
Teachstone coaching cycle costs	5 cycles for each of 115 teachers	8 cycles for each of 107 teachers	$807[b] per cycle	$463,853	$690,535
Principal Supervisory time[c]	16 hours for each of 36 principals	16 hours for each of 34 principals	$109[a] per hour	$62,669	$59,187
Professional Development coordinator time[d]	16 hours for each of 14 PD coordinators	16 hours for each of 14 PD coordinators	$55[a] per hour	$12,412	$12,412
Materials and equipment					
Teachstone video recording kit	One kit for each of 115 teachers	One kit for each of 107 teachers	$0.34[a,f] per teacher	$39	$36
Laptop[e]	115 laptops, each used for 40 minutes per cycle	107 laptops, each used for 40 minutes per cycle	$0.40[a,f] per laptop	$46	$43
Facilities					
Building space for orientation	15 square feet for each of 115 teachers	15 square feet for each of 107 teachers	$0.07[a,f] per square foot	$121	$112
Total cost				$664,454	$904,170
Number of students				2,915	2,701
Cost per student				$228	$335
Cost per student per standard deviation of student achievement				$2,726	$17,382

Note: Number of teachers reflects the number of teachers who received the coaching.

[a] Price data are from the Center for Benefit-Cost Studies of Education.

[b] Price data were provided by Teachstone.

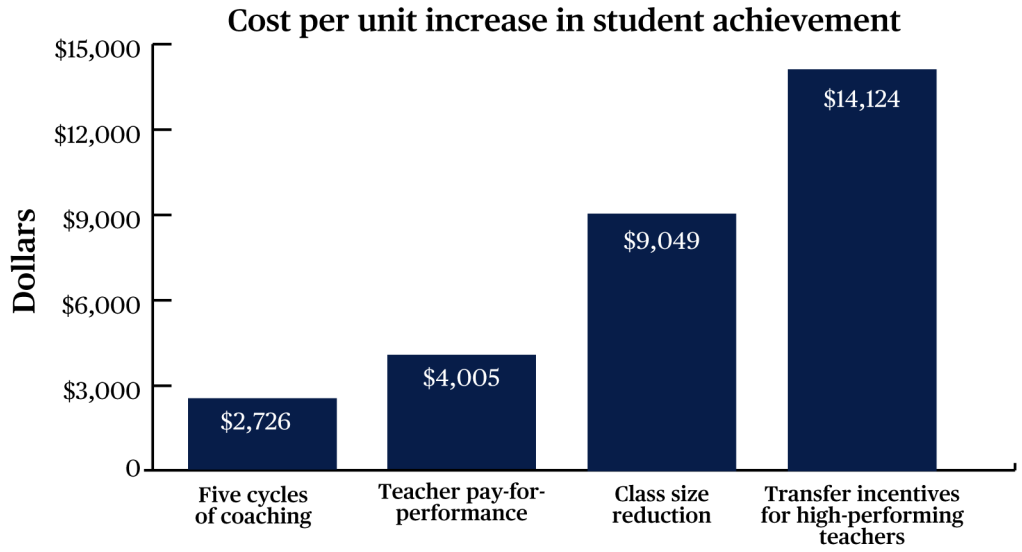
[c] Principals were not directly involved in the coaching but they helped to coordinate video recordings and other activities.

[d] Each district's professional development coordinator was not directly involved in the coaching but served as a point of contact for any implementation issues.

[e] Teachers need a laptop or computer to access their videos and written feedback through the coaching provider's online system, and to hold video conferences with coaches.

[f] The unit price of each laptop, recording kit, and square foot of building space is scaled to reflect the percentage of time the asset is used over the lifetime of the asset.

Exhibit B.20. Comparison of the costs of five cycles of coaching to the costs of other interventions that have been shown to improve student achievement



Source: Administrative student records for the 2017–2018 and 2018–2019 school years.

Note: Exhibit shows cost per standard deviation increase in student achievement. All costs were adjusted for inflation and are expressed in 2018 dollars.

APPENDIX C. SUPPLEMENTAL EXHIBITS AND INFORMATION ON STUDY FINDINGS

This appendix supplements the findings presented in the report. It includes more details on findings in the report, supplemental sensitivity analyses, supplemental information for systematic reviews, and information on realized minimum detectable effects.

C.1 Additional details on findings in the report

This section includes additional information on (1) the effects of the study's coaching on student achievement and (2) teachers' experiences with the coaching and its effects on their teaching practices.

C.1.1 Effects on student achievement

Exhibits 3 and 5 in the report show the effects of the coaching on students' English language arts and math achievement for schools in the five- and eight-cycle coaching groups. Students whose teachers were in the five-cycle coaching group had higher English language arts test scores than students whose teachers did not receive the coaching. Students whose teachers were in the five-cycle group also had higher math scores than students whose teachers did not receive the coaching, but this difference was not statistically significant at the 5 percent level, with a *p*-value of 0.07. Students in the eight-cycle group had similar math and English language arts test scores as students whose teachers did not receive the coaching. Exhibit C.1 presents the effects of the coaching on student achievement and the corresponding *p*-values. Effects are shown in z-score units—they were converted to percentiles in Exhibits 3 and 5 of the report for ease of interpretation, as described in Appendix B.[18]

As discussed in the main body of the report, the coaching may have been particularly effective for novice teachers and teachers with weaker classroom practices at the start of the study. Exhibit 4 in the report shows that, among novice teachers (those in their first five years of teaching) and those with weaker classroom practices at the start of the study, five cycles of coaching led to higher student achievement in both English language arts and math. As shown in Exhibit 6 in the report, eight cycles of coaching did not improve math or English language arts test scores among students of novice teachers or those with weaker classroom practices at the start of the study. As noted in the main body of the report, average scores in both math and English language arts were higher for students taught by novice teachers in the eight-cycle group than for students taught by teachers who did not receive the coaching, but the estimated effects were not statistically significant at the 5 percent level, with *p*-values of 0.08 and 0.11, respectively. Exhibits C.2 and C.3 show these estimated effects and their corresponding *p*-values.

Exhibit C.1. Effects of the study-provided coaching on student achievement

Student achievement (z-score units)	Means			Effects					
	Control group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group		Eight-cycle coaching group		Difference in effects of five and eight cycles	
				Effect	*p*-value	Effect	*p*-value	Difference	*p*-value
Math	-0.20	-0.14	-0.19	0.07	0.07	0.02	0.66	0.05	0.14
English language arts	-0.17	-0.09	-0.16	0.08*	0.01	0.02	0.53	0.06*	0.03
Number of students	**3,034–3,058**	**2,643–2,659**	**2,475–2,495**						
Number of teachers	**102–110**	**85**	**78-82**						
Number of schools	**37**	**36**	**34**						

Source: Administrative student records for the 2017–2018 and 2018–2019 school years.

Note: Test scores were converted to z-scores by subtracting the mean and dividing by the standard deviation of scores for all students in that state and grade level. The study estimated the effects for the five- and eight-cycle coaching groups by comparing outcomes for each of those groups to outcomes for the control group. Sample sizes vary due to the availability of outcome data.

* Statistically significant at the .05 level, two-tailed test.

Exhibit C.2. Effects of the study-provided coaching on student achievement, by teacher experience

Student achievement (z-score units)	Novice teachers					Experienced teachers				
	Control	Five-cycle coaching group		Eight-cycle coaching group		Control	Five-cycle coaching group		Eight-cycle coaching group	
	Mean	Effect	*p*-value	Effect	*p*-value	Mean	Effect	*p*-value	Effect	*p*-value
Math	-0.29	0.14*	0.01	0.13	0.08	-0.17	0.04	0.37	-0.02	0.57
English language arts	-0.30	0.11*	0.01	0.09	0.11	-0.13	0.08*#	0.03	0.00#	0.99
Number of students	**748–762**	**749–778**		**674–746**		**2,212–2,309**	**1,833–1,910**		**1,684–1,709**	
Number of teachers	**30–34**	**23**		**26**		**72–74**	**61–62**		**50–54**	
Number of schools	**19**	**17-18**		**17**		**34**	**29–31**		**30–31**	

Source: Administrative student records for the 2017–2018 and 2018–2019 school years.

Note: Test scores were converted to z-scores by subtracting the mean and dividing by the standard deviation of scores for all students in that state and grade level. The study estimated the effects for the five- and eight-cycle coaching groups by comparing outcomes for each of those groups to outcomes for the control group. Novice teachers are those who have been teaching for five years or less; experienced teachers are those who have been teaching for more than five years. Differences in impacts between novice and experienced teachers were not statistically significant at the .05 level, two-tailed test. Sample sizes vary due to the availability of outcome data.

* Statistically significant at the .05 level, two-tailed test.

Statistically significant difference in impacts between five- and eight-cycle groups at the .05 level, two-tailed test.

Exhibit C.3. Effects of the study-provided coaching on student achievement, by quality of teachers' practices at baseline

	Teachers with weaker teaching practices at baseline					Teachers with stronger teaching practices at baseline				
	Control	Five-cycle coaching group		Eight-cycle coaching group		Control	Five-cycle coaching group		Eight-cycle coaching group	
Outcome	**Means**	**Effect**	***p*-value**	**Effect**	***p*-value**	**Means**	**Effect**	***p*-value**	**Effect**	***p*-value**
Math	-0.46	0.11*#	0.03	-0.05 #	0.41	0.03	0.04	0.40	-0.06 - 0.06	0.30
English language arts	-0.39	0.17*#	0.00	0.05†#	0.19	-0.03	0.09*#	0.03	-0.04 †#	0.26
Number of students	**858–859**	**817–932**		**775–815**		**871–1,115**	**811–897**		**708–734**	
Number of teachers	**31–32**	**24–29**		**23–24**		**28–37**	**24–25**		**24–25**	
Number of schools	**20–21**	**17–20**		**21**		**21–24**	**19–21**		**20–21**	

Source: Administrative student records for the 2017–2018 and 2018–2019 school years.

Note: Test scores were converted to z-scores by subtracting the mean and dividing by the standard deviation of scores for all students in that state and grade level. The study estimated the effects for the five- and eight-cycle coaching groups by comparing outcomes for each of those groups to outcomes for the control group. Teachers with weaker classroom practices at baseline are those who score in the bottom third of the sample; teachers with stronger classroom practices at baseline are those who scored in the top third. Sample sizes vary due to the availability of outcome data.

* Statistically significant at the .05 level, two-tailed test.

Statistically significant difference in impacts between five- and eight-cycle groups at the .05 level, two-tailed test.

† Statistically significant difference in impacts between teachers with weaker and stronger practices at baseline at the .05 level, two-tailed test.

C.1.2 Teachers' experiences with and perceptions of the coaching and its effects on their teaching practices

As discussed in the report, the coaching changed the nature of the feedback teachers received. Exhibit 7 in the report shows that teachers in the coaching groups were more likely to receive feedback that focused on a clearly defined set of teaching practices, provided specific strategies to implement in their classrooms, identified positive aspects of their teaching, and asked questions that encouraged them to reflect on their teaching. Exhibit C.4 presents information on these and other characteristics of the feedback teachers reported receiving.

Similarly, Exhibit 8 in the report shows that about 90 percent of teachers in both coaching groups were more reflective about their teaching as a result of the feedback they received, compared with only 57 percent of teachers who did not receive the coaching. Exhibit 8 also shows that more than 80 percent of teachers in both coaching groups said they made a specific change to their teaching as a result of feedback they received, compared with only 51 percent of teachers who did not receive the coaching. Exhibit C.5 presents this information and other differences in teachers' perceptions of the feedback they received based on observations across the three groups, along with the corresponding *p*-values.

The report notes that the coaching increased the proportion of teachers who received feedback on the aspects of teaching targeted by the coaching. Exhibit C.6 presents this information and the associated *p*-values. The report also notes that about 80 percent of teachers in both coaching groups said they identified aspects of their teaching they needed to improve as a result of watching videos of their own teaching. Exhibit C.7 presents this and other information on the number of video clips viewed by teachers and their reported development from watching the clips.

Exhibit 9 in the report shows that five cycles of coaching did not affect teachers' overall classroom practice score (based on the CLASS rubric), and eight cycles of coaching lowered scores by 0.19 points on a 7-point scale. Exhibit 9 also shows that the coaching did not affect teachers' subscores on practices related to building students' understanding of content. Exhibit C.8 presents the effects of the coaching on teachers' overall and domain-level scores on the CLASS rubric and the corresponding *p*-values.[19]

The study also examined how the coaching's effects on teachers' practices measured by the CLASS rubric differed for novice and experienced teachers and for teachers with weaker and stronger classroom practices at the start of the study. Effects were generally similar for novice and experienced teachers (Exhibit C.9). However, the coaching's negative effects on teachers' scores on the CLASS rubric were particularly pronounced for teachers with stronger teaching practices at the start of the school year (Exhibit C.10).

As described in an endnote of the report, because the coaching improved student achievement in English language arts, the study also examined whether the coaching affected teaching practices specific to English language arts, as measured by the Protocol for Language Arts Teaching Observations (PLATO) rubric. Exhibit C.11 shows that the coaching did not affect teachers' overall scores or subscores for these English language arts–focused practices and presents the corresponding *p*-values.

As noted above, the coaching did not have an effect on teachers' overall practices and had a negative effect on classroom management for teachers who received eight cycles of coaching. The study team examined whether the coaching's effects on practices were related to its effects on student achievement. Exhibit C.12 shows correlations between effects on teachers' practices and effects on student achievement, with all effects estimates at the random assignment block level. Only two of the 60 correlations examined were statistically significant. The lack of strong and consistent patterns of correlations suggests there was not a meaningful relationship between the coaching's effects on teachers' practices and its effects on student achievement.

Exhibit C.4. Characteristics of the feedback teachers received based on observations

Percentage[a] reporting the feedback:	Means			Differences					
	Control group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle vs. control		Eight-cycle vs. control		Five-cycle vs. eight-cycle	
				Difference	*p*-value	Difference	*p*-value	Difference	*p*-value
Examined their performance on a clearly defined set of teaching practices	57	85	87	28*	0.00	30*	0.00	-1	0.77
Provided a score or rating of their performance based on a classroom observation rubric or instrument	48	57	42	9	0.12	-5	0.43	15*	0.02
Provided specific techniques or strategies they could implement in the classroom	36	73	70	36*	0.00	34*	0.00	2	0.72
Referred to specific moments of teaching from their classroom observation	54	91	89	36*	0.00	35*	0.00	1	0.76
Provided questions that encouraged them to reflect on their own teaching	39	77	74	38*	0.00	34*	0.00	3	0.57
Identified aspects of their teaching where they were performing well	53	87	88	34*	0.00	35*	0.00	0	0.90
Identified aspects of their teaching where they needed to improve	39	57	43	17*	0.00	4	0.53	13*	0.02
Included a plan with next steps for them to improve their teaching	25	58	52	32*	0.00	27*	0.00	5	0.39
Involved watching a video of their instruction while discussing feedback	<2	57	62	>55*	0.00	>60*	0.00	-4	0.48
Provided or recommended videos of expert teachers to illustrate practices described in the feedback	3	56	68	53*	0.00	65*	0.00	-12	0.11

Percentage[a] reporting the feedback:	Means			Differences					
	Control group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle vs. control		Eight-cycle vs. control		Five-cycle vs. eight-cycle	
				Difference	*p*-value	Difference	*p*-value	Difference	*p*-value
Provided opportunities for them to observe a demonstration of specific teaching techniques or strategies by the person providing feedback	7	30	47	22*	0.00	39*	0.00	-16*	0.03
Provided an opportunity for them to demonstrate specific teaching techniques or strategies for the person providing feedback	14	54	56	39*	0.00	41*	0.00	-2	0.76
Provided useful or actionable feedback	39	77	74	38*	0.00	35*	0.00	2	0.69
Number of teachers	**128–130**	**108–110**	**102–104**						
Number of schools	**37**	**36**	**34**						

Source: Teacher survey administered in spring 2019.

Note: Values indicate the percentage of teachers who reported the feedback included the specific content "most of the time" or "always." A < or > indicates that the exact percentage has been withheld to protect respondent confidentiality in accordance with National Center for Education Statistics statistical standards, but the percentage is less than or greater than the number following the < or > symbol. Sample sizes vary due to the availability of outcome data.

[a] Differences between groups may differ from differences in reported means due to rounding.

* Statistically significant at the .05 level, two-tailed test.

Exhibit C.5. Teachers' perceptions of feedback they received based on observations

	Means			Differences					
				Five-cycle vs. control		Eight-cycle vs. control		Five-cycle vs. eight-cycle	
Percentage reporting[a]	Control group	Five-cycle coaching group	Eight-cycle coaching group	Difference	*p*-value	Difference	*p*-value	Difference	*p*-value
Received feedback that was easy to understand	60	95	93	34*	0.00	33*	0.00	1	0.68
Received feedback that provided specific ideas about how they could improve their performance	50	87	90	36*	0.00	40*	0.00	-3	0.48
Feedback made them more reflective about their teaching	57	89	90	32*	0.00	33*	0.00	-1	0.80
Believe in the long run that students will benefit from the feedback they received	59	88	86	28*	0.00	26*	0.00	1	0.77
Made a specific change to their teaching as a result of the feedback	51	81	87	29*	0.00	35*	0.00	-6	0.30
Number of teachers	**129–130**	**108–110**	**103–104**						
Number of schools	**37**	**36**	**34**						

Source: Teacher survey administered in spring 2019.

Note: Values indicate the percentage of teachers who reported that they received feedback based on in-person or video-based observations and "agree somewhat" or "agree strongly." Sample sizes vary due to the availability of outcome data.

[a] Differences between groups may differ from differences in reported means due to rounding.

* Statistically significant at the .05 level, two-tailed test.

Exhibit C.6. Focus of feedback teachers received based on observations

Percentage[a] reporting that they received feedback focused on:	Means			Differences					
				Five-cycle vs. control		Eight-cycle vs. control		Five-cycle vs. eight-cycle	
	Control group	Five-cycle coaching group	Eight-cycle coaching group	Difference	*p*-value	Difference	*p*-value	Difference	*p*-value
Managing student behavior	23	37	42	13*	0.03	18*	0.00	-5	0.42
Managing instructional time and routines	27	46	61	19*	0.00	33*	0.00	-14*	0.02
Engaging students in classroom instruction through clear and interesting lessons and materials	35	73	73	38*	0.00	37*	0.00	0	0.94
Providing feedback that extends students' learning and encourages their participation	33	74	80	41*	0.00	47*	0.00	-5	0.38
Leading discussions that build a deeper understanding of the content	29	76	79	47*	0.00	50*	0.00	-3	0.64
Supporting students' use of higher-level thinking skills	37	74	81	37*	0.00	44*	0.00	-7	0.25
Responding to the academic, social, and emotional needs of individual students and the entire class	18	47	58	28*	0.00	39*	0.00	-11	0.08
Developing lesson plans that are aligned to learning goals and include engaging activities	29	48	62	18*	0.01	33*	0.00	-14	0.06
Establishing an environment where the teacher and the teacher's students support and respect each other	25	49	62	23*	0.00	36*	0.00	-12	0.08
Incorporating students' perspectives and interests into classroom activities	16	52	59	35*	0.00	42*	0.00	-6	0.29

Percentage[a] reporting that they received feedback focused on:	Means			Differences					
				Five-cycle vs. control		Eight-cycle vs. control		Five-cycle vs. eight-cycle	
	Control group	Five-cycle coaching group	Eight-cycle coaching group	Difference	*p*-value	Difference	*p*-value	Difference	*p*-value
Building students' understanding of core academic content	25	59	67	33*	0.00	42*	0.00	-8	0.24
Number of teachers	**129–130**	**107–110**	**102–104**						
Number of schools	**36–37**	**36**	**34**						

Source: Teacher survey administered in spring 2019.

Note: Values indicate the percentage of teachers who reported that they received feedback based on in-person or video-based observations focused on specific areas "to a moderate extent" or "to a great extent." Sample sizes vary due to the availability of outcome data.

[a] Differences between groups may differ from differences in reported means due to rounding.

* Statistically significant at the .05 level, two-tailed test.

Exhibit C.7. Average number of video clips viewed by teachers and reported development from watching video clips

	Five-cycle coaching group	Eight-cycle coaching group
Video clips viewed		
Average number of video clips of their own teaching that teachers viewed during each coaching cycle	2.6	2.6
Percentage of coaching cycles during which teachers viewed all of the clips of their own teaching	78	77
Average number of exemplar videos teachers viewed during each coaching cycle	1.00	0.80
Percentage of coaching cycles during which teachers viewed exemplar videos	34	32
Teachers' reported development from watching video clips		
Percentage of teachers reporting that they:		
Identified aspects of their teaching that needed to improve as a result of watching video clips of their teaching	81	84
Made a specific change to their teaching based on something they saw in a video clip of their teaching	79	79
Learned something about their own teaching practice by watching video clips of their teaching	85	84
Noticed student behaviors or reactions that they had not previously noticed while teaching after watching video clips of their teaching	79	81
Number of teachers	**105–111**	**102–103**
Number of schools	**34–36**	**32–34**

Source: Data collected from Teachstone, 2018–2019 school year and teacher survey administered in spring 2019.

Note: Teachers were expected to view three clips of their own teaching for each coaching cycle and two exemplar clips. The averages include coaching cycles in which teachers did not view a clip of their own teaching or did not view an exemplar clip. Sample sizes vary due to the availability of outcome data.

Exhibit C.8. Effects of the study-provided coaching on teachers' general classroom practices

Outcome	Means			Effects					
	Control group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group		Eight-cycle coaching group		Difference in effects of five and eight cycles	
				Effect	*p*-value	Effect	*p*-value	Difference	*p*-value
Overall CLASS score	4.59	4.46	4.41	-0.13	0.12	-0.18*	0.03	0.050	0.52
Classroom management	6.45	6.28	6.19	-0.18*	0.01	-0.26*	0.00	0.080	0.23
Building students' understanding	3.57	3.43	3.40	-0.14	0.22	-0.17	0.11	0.030	0.76
Building supportive relationships with students	4.18	4.08	4.01	-0.10	0.39	-0.17	0.14	0.070	0.56
Building students' engagement	5.37	5.25	5.21	-0.12	0.24	-0.15	0.14	0.040	0.66
Number of teachers	**125**	**99**	**98**						
Number of schools	**36**	**33**	**31**						

Source: CLASS ratings of video-recorded classroom observations in spring 2019.

Note: The study estimated the effects for the five- and eight-cycle coaching groups by comparing outcomes for each of those groups to outcomes for the control group. The overall CLASS score and domain scores range from 1 to 7, with higher values indicating more positive outcomes. Each domain score is the average of scores on a series of dimensions. The classroom management domain includes dimensions for behavior management, productivity, and negative climate. The building students' understanding domain includes dimensions for instructional learning formats, content understanding, analysis and inquiry, quality of feedback, and instructional dialogue. The building supportive relationships with students domain includes dimensions for positive climate, teacher sensitivity, and regard for adolescent perspectives. The overall CLASS score is an average of 12 dimension-level scores—those that make up the three domains along with an additional stand-alone dimension for building student engagement, which also ranges from 1 to 7.

* Statistically significant at the .05 level, two-tailed test.

CLASS = Classroom Assessment Scoring System.

Exhibit C.9. Effects of the study-provided coaching on teachers' general classroom practices, by teacher experience

	Novice teachers					Experienced teachers				
	Control	Five-cycle coaching group		Eight-cycle coaching group		Control	Five-cycle coaching group		Eight-cycle coaching group	
Outcome	**Means**	**Effect**	***p*-value**	**Effect**	***p*-value**	**Means**	**Effect**	***p*-value**	**Effect**	***p*-value**
Overall CLASS score	4.50	-0.12	0.45	-0.30*	0.02	4.62	-0.13	0.15	-0.10	0.29
Classroom management	6.47	-0.15#	0.17	-0.39*#	0.00	6.45	-0.19*	0.01	-0.20*	0.02
Building students' understanding	3.40	-0.03	0.88	-0.21	0.14	3.62	-0.16	0.17	-0.10	0.42
Building supportive relationships with students	4.12	-0.25	0.29	-0.42*	0.04	4.20	-0.04	0.78	-0.04	0.75
Building students' engagement	5.29	-0.03	0.86	-0.11	0.53	5.39	-0.14	0.25	-0.15	0.24
Number of teachers	**34**	**32**		**33**		**91**	**67**		**65**	
Number of schools	**19**	**21**		**21**		**35**	**29**		**30**	

Source: CLASS ratings of video-recorded classroom observations in spring 2019, teacher participation forms administered in fall 2018.

Note: The study estimated the effects for the five- and eight-cycle coaching groups by comparing outcomes for each of those groups to outcomes for the control group. The overall CLASS score and domain scores range from 1 to 7, with higher values indicating more positive outcomes. Each domain score is the average of scores on a series of dimensions. The classroom management domain includes dimensions for behavior management, productivity, and negative climate. The building students' understanding domain includes dimensions for instructional learning formats, content understanding, analysis and inquiry, quality of feedback, and instructional dialogue. The building supportive relationships with students domain includes dimensions for positive climate, teacher sensitivity, and regard for adolescent perspectives. The overall CLASS score is an average of 12 dimension-level scores—those that make up the three domains along with an additional stand-alone dimension for building student engagement, which also ranges from 1 to 7. Novice teachers are those who have been teaching for five years or less; experienced teachers are those who have been teaching for more than five years. Differences in impacts between novice and experienced teachers were not statistically significant at the .05 level, two-tailed test.

* Statistically significant at the .05 level, two-tailed test.

Statistically significant difference in impacts between five- and eight-cycle groups at the .05 level, two-tailed test.

CLASS = Classroom Assessment Scoring System.

Exhibit C.10. Effects of the study-provided coaching on teachers' general classroom practices, by quality of teachers' practices at baseline

	Teachers with weaker teaching practices at baseline					Teachers with stronger teaching practices at baseline				
	Control	Five-cycle coaching group		Eight-cycle coaching group		Control	Five-cycle coaching group		Eight-cycle coaching group	
Outcome	**Means**	**Effect**	***p*-value**	**Effect**	***p*-value**	**Means**	**Effect**	***p*-value**	**Effect**	***p*-value**
Overall CLASS score	4.19	0.24†	0.15	0.04	0.78	4.85	-0.36*†	0.01	-0.30*	0.04
Classroom management	6.22	0.01	0.91	-0.10	0.51	6.55	-0.19	0.12	-0.25*	0.05
Building students' understanding	3.09	0.32†	0.10	0.11	0.51	3.87	-0.46*†	0.00	-0.30	0.08
Building supportive relationships with students	3.70	0.30†	0.18	0.06	0.78	4.54	-0.32†	0.17	-0.33	0.14
Building students' engagement	5.03	0.27†	0.21	0.17	0.40	5.63	-0.31†	0.06	-0.36*	0.04
Number of teachers	**36**	**37**		**33**		**44**	**33**		**29**	
Number of schools	**24**	**23**		**23**		**26**	**22**		**22**	

Source: CLASS ratings of video-recorded classroom observations in fall 2018 and spring 2019.

Note: The study estimated the effects for the five- and eight-cycle coaching groups by comparing outcomes for each of those groups to outcomes for the control group. The overall CLASS score and domain scores range from 1 to 7, with higher values indicating more positive outcomes. Each domain score is the average of scores on a series of dimensions. The classroom management domain includes dimensions for behavior management, productivity, and negative climate. The building students' understanding domain includes dimensions for instructional learning formats, content understanding, analysis and inquiry, quality of feedback, and instructional dialogue. The building supportive relationships with students domain includes dimensions for positive climate, teacher sensitivity, and regard for adolescent perspectives. The overall CLASS score is an average of 12 dimension-level scores—those that make up the three domains along with an additional stand-alone dimension for building student engagement, which also ranges from 1 to 7. Teachers with weaker classroom practices at baseline are those who score in the bottom third of the sample; teachers with stronger classroom practices at baseline are those who scored in the top third. Differences in impacts between teachers with weaker and stronger practices at baseline are not statistically significant at the .05 level, two-tailed test.

* Statistically significant at the .05 level, two-tailed test.

† Statistically significant difference in impacts between teachers with weaker and stronger practices at baseline at the .05 level, two-tailed test.

CLASS = Classroom Assessment Scoring System.

Exhibit C.11. Effects of the study-provided coaching on teachers' English language arts-specific practices

Outcome	Means			Effects					
	Control group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group		Eight-cycle coaching group		Difference in effects of five and eight cycles	
				Effect	*p*-value	Effect	*p*-value	Difference	*p*-value
Overall PLATO score	2.35	2.31	2.29	-0.03	0.59	-0.06	0.36	0.020	0.70
Instruction and scaffolding	1.87	1.85	1.77	-0.03	0.70	-0.10	0.12	0.080	0.28
Disciplinary demand	2.14	2.18	2.23	0.04	0.67	0.10	0.32	-0.060	0.54
Representations and use of content	2.26	2.18	2.15	-0.08	0.28	-0.11	0.16	0.030	0.71
Classroom environment	3.72	3.64	3.60	-0.08	0.17	-0.12*	0.04	0.040	0.54
Number of teachers	**96**	**71**	**70**						
Number of schools	**36**	**32**	**31**						

Source: PLATO ratings of video-recorded English language arts classroom observations in spring 2019.

Note: The study estimated the effects for the five- and eight-cycle coaching groups by comparing outcomes for each of those groups to outcomes for the control group. The overall PLATO score and domain scores range from 1 to 4, with higher values indicating more positive outcomes. Each domain score is the average of scores on a series of elements. The instruction and scaffolding domain includes elements for modeling and use of models, strategy use and instruction, feedback, and accommodations for language learning. The disciplinary demand domain includes elements for intellectual challenge, classroom discourse, and text-based instruction. The representations and use of content domain includes elements for representation of content, connections to prior academic knowledge, and purpose. The classroom environment domain includes elements for behavior management and time management. The overall PLATO score is an average of the 13 elements that make up the four domains. Differences in impacts between teachers in the five-cycle and eight-cycle groups are not statistically significant at the .05 level, two-tailed test. The effect of the study's coaching on the representations and use of content and classroom environment domains should be interpreted with caution given the low internal consistency reliability shown in Exhibit B.12.

* Statistically significant at the .05 level, two-tailed test.

PLATO = Protocol for Language Arts Teaching Observations.

Exhibit C.12. Correlations between the study-provided coaching's effects on teachers' classroom practices and its effects on student achievement

Teachers' classroom practices	Student test scores			
	Five-cycle group		Eight-cycle group	
	Math correlation (*p*-value)	English language arts correlation (*p*-value)	Math correlation (*p*-value)	English language arts correlation (*p*-value)
Overall CLASS score	-0.08 (0.68)	-0.03 (0.90)	-0.05 (0.79)	-0.08 (0.69)
Classroom management	0.12 (0.53)	0.00 (0.99)	-0.03 (0.90)	-0.07 (0.71)
Building students' understanding	-0.14 (0.49)	-0.04 (0.83)	-0.08 (0.70)	-0.11 (0.59)
Building supportive relationships with students	-0.07 (0.71)	0.01 (0.96)	-0.02 (0.91)	0.01 (0.97)
Building students' engagement	-0.03 (0.86)	-0.07 (0.72)	0.01 (0.97)	-0.13 (0.52)
Correlation among novice teachers				
Overall CLASS score	0.20 (0.35)	-0.05 (0.82)	0.06 (0.78)	-0.03 (0.90)
Classroom management	0.41 (0.05)*	0.08 (0.70)	0.04 (0.86)	-0.10 (0.65)
Building students' understanding	0.17 (0.42)	-0.02 (0.94)	0.01 (0.97)	0.10 (0.64)
Building supportive relationships with students	-0.07 (0.75)	-0.19 (0.39)	0.16 (0.47)	-0.09 (0.69)
Building students' engagement	0.33 (0.12)	0.07 (0.76)	-0.19 (0.37)	-0.09 (0.67)
Correlation among teachers with weaker practices at baseline				
Overall CLASS score	0.19 (0.35)	-0.01 (0.97)	0.23 (0.25)	-0.03 (0.88)
Classroom management	0.41 (0.03)*	0.32 (0.12)	0.53 (0.00)*	0.11 (0.62)
Building students' understanding	0.09 (0.67)	-0.08 (0.71)	0.01 (0.96)	-0.09 (0.68)
Building supportive relationships with students	0.06 (0.77)	-0.20 (0.35)	0.13 (0.53)	-0.07 (0.73)
Building students' engagement	0.29 (0.14)	0.23 (0.27)	0.35 (0.07)	0.08 (0.69)
Number of random assignment blocks	**24–28**	**23–28**	**24–28**	**23–28**

Source: CLASS ratings of video-recorded classroom observations in spring 2019 and administrative student records for 2017–2018 and 2018–2019 school years.

Note: Effects on CLASS scores and student test scores are standardized into z-scores by subtracting the mean and dividing by the standard deviation. Novice teachers are those who have been teaching for five years or less; experienced teachers are those who have been teaching for more than five years. Quality of teachers' teaching practices is defined based on teachers' baseline CLASS scores. The CLASS ranges from 1 to 7. Teachers with weaker teaching practices at the start of the study are those who scored in the bottom third of CLASS scores for the sample. Sample sizes vary due to the availability of outcome data.

* Correlations between block-level effects are statistically significant at the .05 level.

CLASS = Classroom Assessment Scoring System.

C.2 Supplemental sensitivity analyses

This section includes supplemental sensitivity analyses, examining implementation of the coaching for different groups of teachers and how the coaching's effects varied across districts, random assignment blocks, and coaches.

Exhibits C.13 and C.14 show that coaches implemented the coaching similarly for teachers in the five- and eight-cycle groups, overall as well as among novice and experienced teachers and teachers with weaker and stronger practices at the start of the study. Coaches generally covered the same aspects of teaching and spent similar amounts of time in coaching conferences across all these groups. However, a key difference for teachers in the five- and eight-cycle groups was the timing of the coaching cycles. Because they had to cover more cycles in the same amount of time, coaches shortened the length of the coaching cycles and completed them later in the school year for teachers in the eight-cycle group. Exhibit C.13 presents details on the focus of the coaching across these groups of teachers, and Exhibit C.14 presents details on key features of the coaching received, including the timing of the coaching cycles.

Exhibits C.15–C.16 show the effects of the coaching on student achievement by district and random assignment block, and Exhibits C.17–C.18 show how they varied by coach. Estimated effects did not vary to a statistically significant degree across districts and random assignment blocks but did vary across coaches.

Exhibit C.19 shows the effects of the coaching on teachers' practices varied by district and random assignment block, and Exhibit C.20 shows how they varied by coach. Estimated effects on teachers' practices also did not vary to a statistically significant degree across districts and random assignment blocks but did vary across coaches.

Exhibit C.13. Average number of cycles focused on each CLASS dimension and domain, by coaching group, teacher experience, and quality of teachers' practices at baseline

Average number of cycles focused on...	All teachers		Novice teachers		Experienced teachers		Teachers with weaker teaching practices at baseline		Teachers with stronger teaching practices at baseline	
	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group
Classroom management	**0.9**	**1.3**	**1.0**	**1.4**	**0.9**	**1.3**	**1.0**	**1.3**	**0.8**	**1.3**
Behavior management	0.3	0.6	0.4	0.8	0.2	0.5	0.4	0.7	0.1	0.5
Productivity	0.7	0.7	0.6	0.6	0.7	0.8	0.6	0.6	0.7	0.7
Negative climate	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Building students' understanding	**3.8**	**6.6**	**3.9**	**6.8**	**3.8**	**6.5**	**3.8**	**6.7**	**3.8**	**6.2**
Instructional learning formats	1.2	1.8	1.2	1.9	1.2	1.8	1.2	1.9	1.2	1.6
Content understanding	1.1	2.2	1.1	2.1	1.1	2.3	1.1	2.3	1.1	2.0
Analysis and inquiry	1.0	2.1	0.9	2.1	1.1	2.0	0.9	2.0	1.1	2.1
Quality of feedback	1.1	1.9	1.0	2.1	1.1	1.8	1.0	1.9	1.1	1.8
Instructional dialogue	0.8	1.8	1.0	1.7	0.8	1.9	0.9	1.9	0.9	1.7
Building supportive relationships with students	**1.0**	**1.0**	**1.0**	**1.1**	**1.0**	**1.0**	**1.0**	**1.1**	**1.0**	**0.9**
Positive climate	0.1	0.1	0.1	0.2	0.0	0.1	0.1	0.1	0.1	0.1
Teacher sensitivity	0.5	0.5	0.4	0.6	0.5	0.4	0.5	0.5	0.5	0.4
Regard for student perspective	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.4	0.4
Student engagement	**0.1**	**0.1**	**0.1**	**0.2**	**0.1**	**0.0**	**0.1**	**0.1**	**0.1**	**0.1**
Number of teachers	**105**	**102**	**32**	**33**	**73**	**69**	**39**	**35**	**35**	**32**

Source: Data collected from Teachstone, 2018–2019 school year.

Note: Teachers with weaker classroom practices at baseline are those who score in the bottom third of the sample; teachers with stronger classroom practices at baseline are those who scored in the top third.

Exhibit C.14. Key features of the coaching received, by coaching group, teacher experience, and quality of teachers' practices at baseline

	All teachers		Novice teachers		Experienced teachers		Teachers with weaker teaching practices at baseline		Teachers with stronger teaching practices at baseline	
	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group
Average number of days per coaching cycle										
All assigned cycles	28	22	28	22	28	22	28	22	28	23
Cycles 1-5	28	25	28	25	28	25	28	26	28	27
Cycles 6-8	n.a.	17	n.a.	18	n.a.	16	n.a.	16	n.a.	17
Average number of teaching practices (CLASS indicators) covered across all coaching cycles										
All assigned cycles	10.0	15.2	9.9	15.2	10.0	15.2	9.9	15.4	9.9	14.7
Cycles 1-5	10.0	10.0	9.9	10.1	10.0	10.0	9.9	10.2	9.9	9.8
Cycles 6-8	n.a.	6.6	n.a.	6.4	n.a.	6.7	n.a.	7.0	n.a.	6.2
Average number of teaching practices (CLASS indicators) covered across all coaching cycles - Classroom organization										
All assigned cycles	1.6	2.1	1.8	2.3	1.5	2.1	1.7	2.2	1.3	2.2
Cycles 1-5	1.6	2.1	1.8	2.3	1.5	2.1	1.7	2.2	1.3	2.2
Cycles 6-8	n.a.	0.0	n.a.	0.0	n.a.	0.0	n.a.	0.0	n.a.	0.0
Average number of teaching practices (CLASS indicators) covered across all coaching cycles - Building supportive relationships with students										
All assigned cycles	1.6	1.7	1.4	1.7	1.6	1.7	1.5	1.7	1.6	1.6
Cycles 1-5	1.6	1.7	1.4	1.7	1.6	1.7	1.5	1.7	1.6	1.6
Cycles 6-8	n.a.	0.0	n.a.	0.1	n.a.	0.0	n.a.	0.0	n.a.	0.0
Average number of teaching practices (CLASS indicators) covered across all coaching cycles - Building students' understanding of content										
All assigned cycles	6.7	11.3	6.7	11.0	6.7	11.4	6.7	11.3	6.8	10.9
Cycles 1-5	6.7	6.2	6.7	6.1	6.7	6.2	6.7	6.2	6.8	6.0
Cycles 6-8	n.a.	6.5	n.a.	6.1	n.a.	6.7	n.a.	7.0	n.a.	6.2

	All teachers		Novice teachers		Experienced teachers		Teachers with weaker teaching practices at baseline		Teachers with stronger teaching practices at baseline	
	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group	Five-cycle coaching group	Eight-cycle coaching group
Proportion of teachers who completed their first cycle in…										
September	7.6	9.8	0.0	18.2	11.0	5.8	< 7.7	< 8.6	< 8.6	12.5
October	64.8	55.9	> 65.6	60.6	< 64.4	53.6	> 59.0	> 45.7	> 68.6	50.0
November	21.9	24.5	25.0	21.2	20.5	26.1	25.6	34.3	14.3	25.0
December or later	5.7	9.8	< 9.4	0.0	> 4.1	14.5	< 7.7	11.4	8.6	12.5
Proportion of teachers who completed their fifth cycle in…										
November or December	5.1	13.3	0.0	15.2	7.4	12.3	< 8.1	8.8	< 9.1	< 10.3
January	26.3	40.8	38.7	48.5	20.6	36.9	27.0	> 35.3	30.3	34.5
February	38.4	32.7	25.8	27.3	44.1	35.4	> 35.1	29.4	> 33.3	> 34.5
March	22.2	10.2	25.8	9.1	20.6	10.8	21.6	17.6	18.2	< 10.3
April	8.1	0.0	9.7	0.0	7.4	0.0	8.1	0.0	< 9.1	0.0
May	0.0	3.1	0.0	0.0	0.0	4.6	0.0	< 8.8	0.0	< 10.3
Proportion of teachers who completed their eighth cycle in…										
January or February	n.a.	17.4	n.a.	22.6	n.a.	14.8	n.a.	12.5	n.a.	11.1
March	n.a.	43.5	n.a.	41.9	n.a.	44.3	n.a.	43.8	n.a.	48.1
April or May	n.a.	39.1	n.a.	35.5	n.a.	41.0	n.a.	43.8	n.a.	40.7
Number of teachers	**99–105**	**92–102**	**31–32**	**31–33**	**68–73**	**61–69**	**37–39**	**32–35**	**33–35**	**27–32**

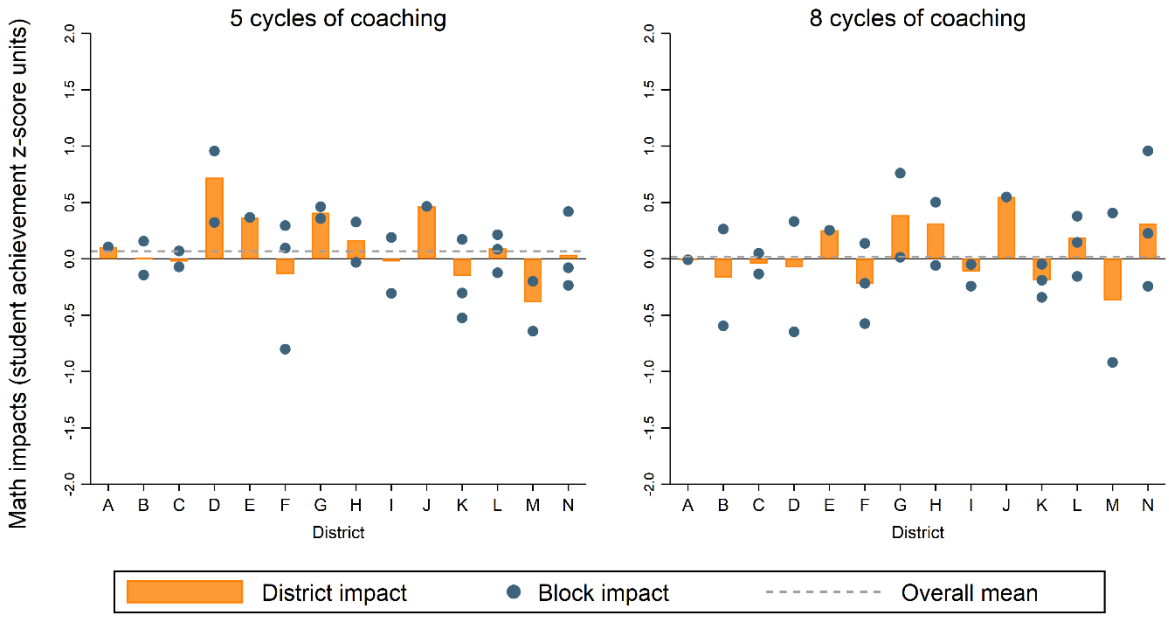
Source: Data collected from Teachstone, 2018–2019 school year.

Notes: To determine how many CLASS indicators coaches covered in each coaching cycle, the study team coded the written summaries that coaches provided to teachers after every coaching cycle. The written summaries describe the specific CLASS indicators covered in the coaching cycle. Teachers with weaker classroom practices at baseline are those who score in the bottom third of the sample; teachers with stronger classroom practices at baseline are those who scored in the top third. A < or > indicates that the exact percentage has been withheld to protect

respondent confidentiality in accordance with National Center for Education Statistics statistical standards, but the percentage is less than or greater than the number following the < or > symbol. Sample includes all teachers randomly assigned to the five- or eight-cycle coaching groups. Sample sizes vary due to the availability of outcome data.

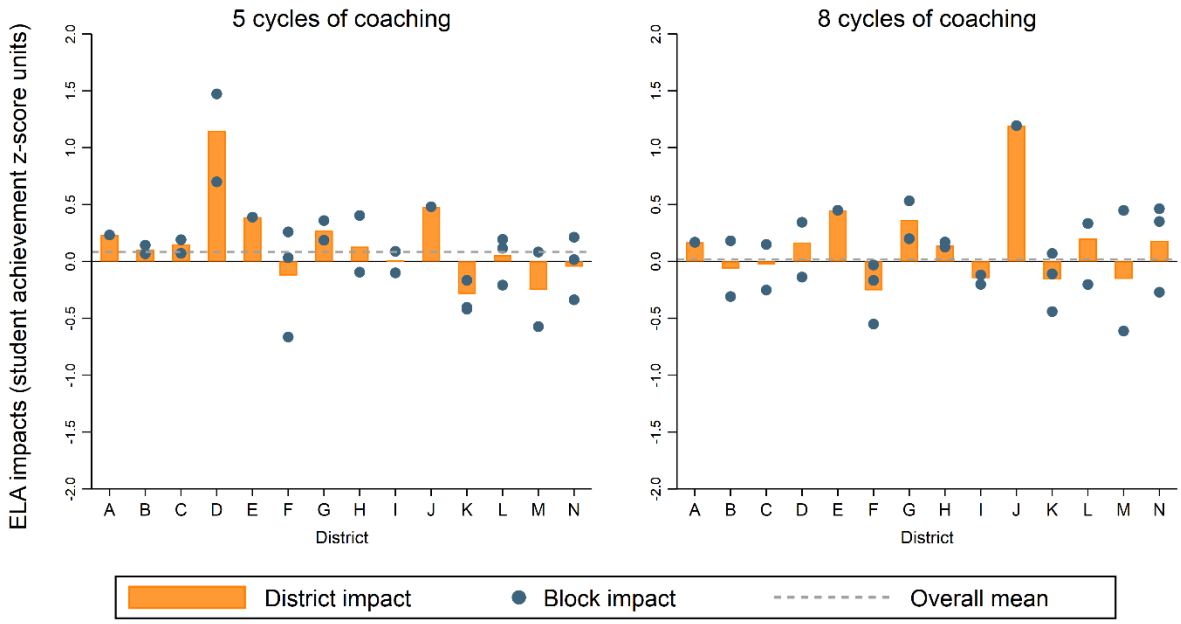
n.a. = not applicable.

Exhibit C.15. Effects of the study-provided coaching on students' math achievement, by district and random assignment block



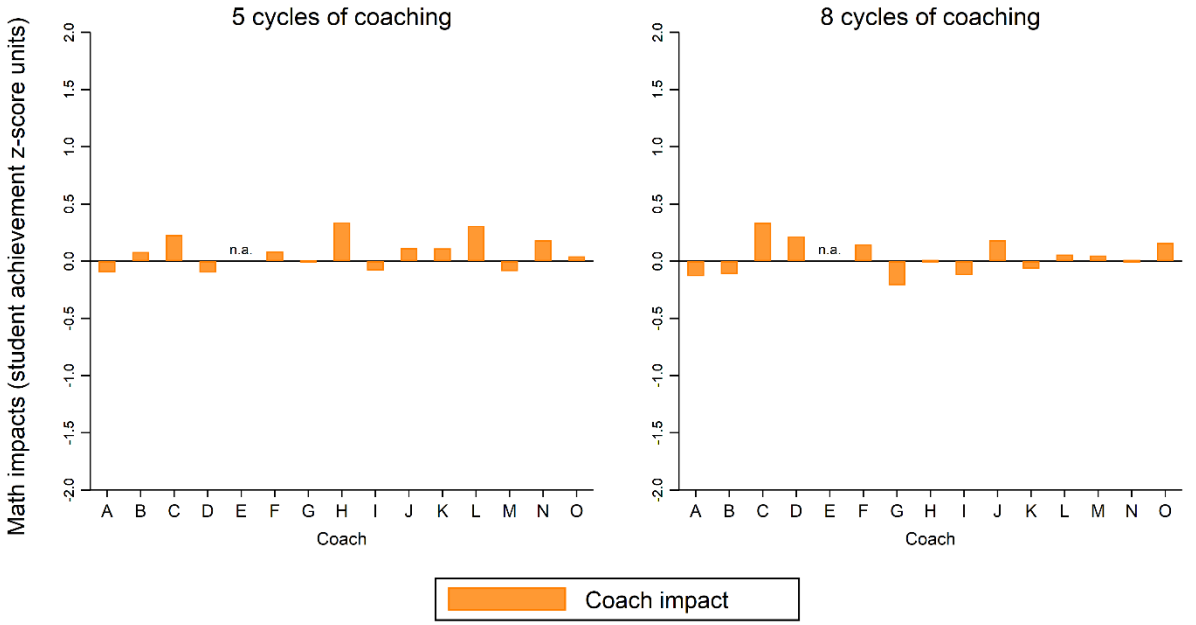
Note: A Monte Carlo permutation test of the null hypothesis that five-cycle effects do not vary across districts has a *p*-value of 0.356. A test of the null hypothesis of equal variance in block-level means between the control group and the group assigned to receive five cycles of coaching has a *p*-value of 0.645. A Monte Carlo permutation test of the null hypothesis that eight-cycle effects do not vary across districts has a *p*-value of 0.764. A test of the null hypothesis of equal variance in block-level means between the control group and the group assigned to receive eight cycles of coaching has a *p*-value of 0.977.

Exhibit C.16. Effects of the study-provided coaching on students' English language arts achievement, by district and random assignment block



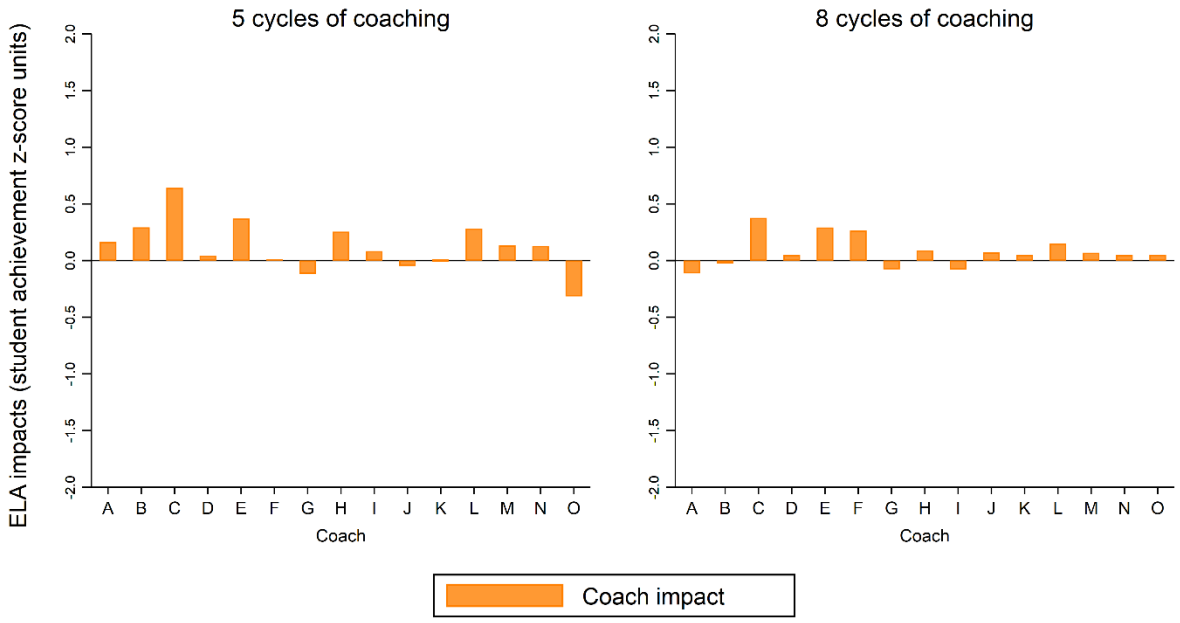
Note: A Monte Carlo permutation test of the null hypothesis that five-cycle effects do not vary across districts has a *p*-value of 0.164. A test of the null hypothesis of equal variance in block-level means between the control group and the group assigned to receive five cycles of coaching has a *p*-value of 0.499. A Monte Carlo permutation test of the null hypothesis that eight-cycle effects do not vary across districts has a *p*-value of 0.082. A test of the null hypothesis of equal variance in block-level means between the control group and the group assigned to receive eight cycles of coaching has a *p*-value of 0.904.

Exhibit C.17. Effects of the study-provided coaching on students' math achievement, by coach



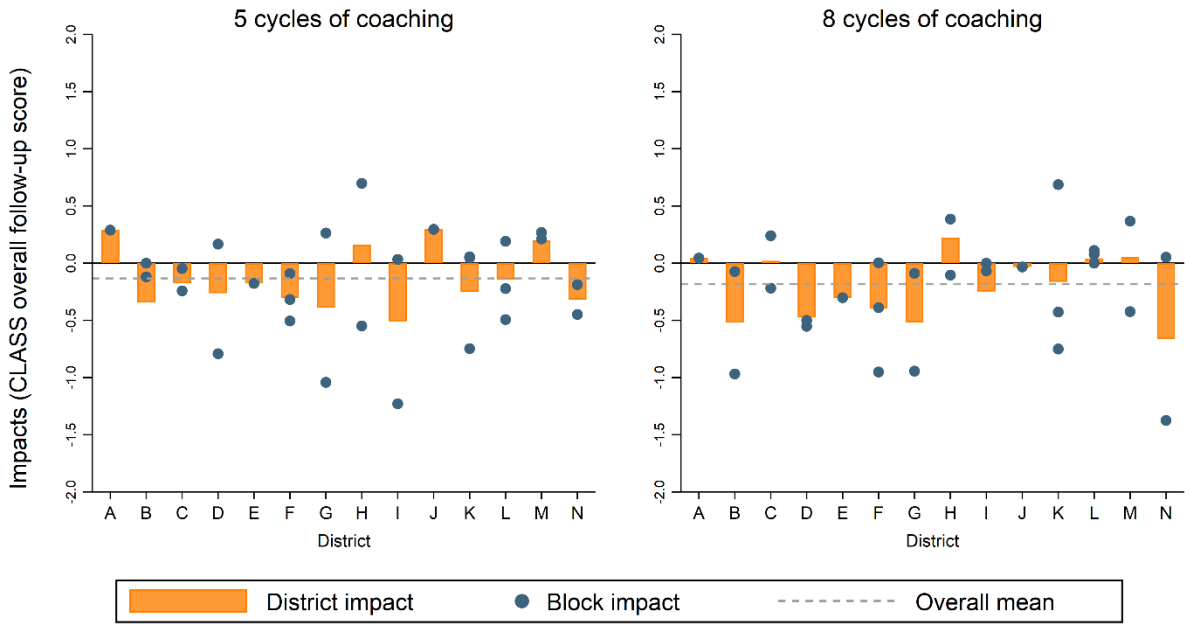
Note: A test of the null hypothesis of equal variance in coach-level means between the control group and the group assigned to receive five cycles of coaching has a *p*-value of 0.00. A test of the null hypothesis of equal variance in coach-level means between the control group and the group assigned to receive eight cycles of coaching has a *p*-value of 0.00.

Exhibit C.18. Effects of the study-provided coaching on students' English language arts achievement, by coach



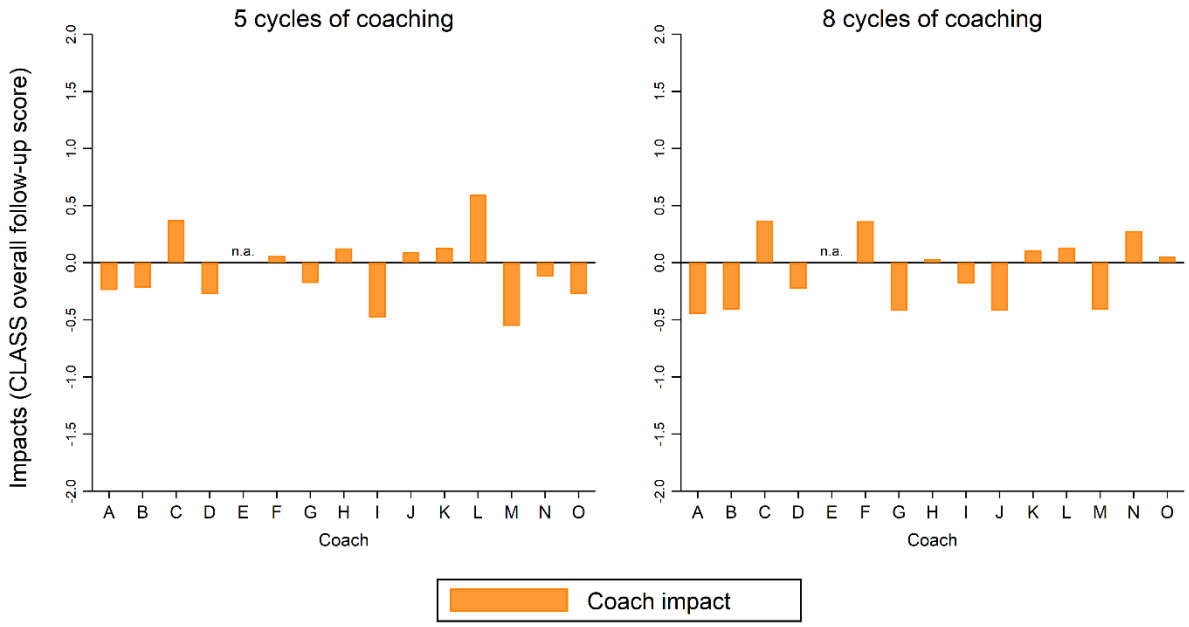
Note: A test of the null hypothesis of equal variance in coach-level means between the control group and the group assigned to receive five cycles of coaching has a *p*-value of 0.00. A test of the null hypothesis of equal variance in coach-level means between the control group and the group assigned to receive eight cycles of coaching has a *p*-value of 0.00.

Exhibit C.19. Effects of the study-provided coaching on teachers' overall CLASS scores, by district and random assignment block



Note: A Monte Carlo permutation test of the null hypothesis that five-cycle effects do not vary across districts has a *p*-value of 0.892. A test of the null hypothesis of equal variance in block-level means between the control group and the group assigned to receive five cycles of coaching has a *p*-value of 0.592. A Monte Carlo permutation test of the null hypothesis that eight-cycle effects do not vary across districts has a *p*-value of 0.921. A test of the null hypothesis of equal variance in block-level means between the control group and the group assigned to receive eight cycles of coaching has a *p*-value of 0.243.

Exhibit C.20. Effects of the study-provided coaching on teachers' overall CLASS scores, by coach



Note: A test of the null hypothesis of equal variance in coach-level means between the control group and the group assigned to receive five cycles of coaching has a *p*-value of 0.01. A test of the null hypothesis of equal variance in coach-level means between the control group and the group assigned to receive eight cycles of coaching has a *p*-value of 0.00.

C.3 Supplemental information for systematic reviews

Systematic reviews of evidence on the effects of educational interventions such as those conducted by the U.S. Department of Education's What Works Clearinghouse (WWC) often require specific types of information to evaluate the quality of a study. This section presents additional information that a systematic review might need to assess the quality of the findings.

According to WWC Version 4.1 Standards, threats to the integrity of a cluster randomized controlled trial are limited if (1) cluster-level attrition (in the case of this study, attrition of schools from the sample) is low, (2) individual nonresponse (of students or teachers, depending on the level of the outcome) is low, and (3) there is no risk of bias due to individuals (students or teachers, depending on the level of the outcome) joining the analytic sample after randomization (WWC 2020). If the study meets the listed conditions, baseline equivalence is not required to meet WWC standards without reservations because the integrity of random assignment ensures the outcomes are not related to any observed or unobserved characteristics other than assignment to the treatment group. If any of the listed conditions are not met, then the study may need to meet the WWC's baseline equivalence requirement by showing that differences between the treatment and control groups on key baseline covariates are smaller than 0.25 standard deviations (the WWC's limit for satisfying baseline equivalence after adjusting for baseline characteristics, which the study does). In general, the WWC focuses on baseline measures of the outcome variables (in this study, student assessment scores and teacher practice measures) when examining baseline equivalence rather than other covariates, such as student demographics.

Students who joined mid-year were not included in the analytic sample. In addition, the study's coaching likely did not affect teachers' decisions to join or leave study schools. However, to confirm that findings are similar with and without teachers who joined mid-year, the section presents information for both the main analytic teacher sample (those in study schools at the time of the end-of-year classroom observations) and an alternative sample of teachers (those who were in study schools in the first six weeks of the school year).

Exhibit C.21 shows unadjusted means, standard deviations, and sample sizes for the main analytic samples of students and teachers, and the alternative sample of teachers. Some teachers and students in the analytic sample are missing baseline information; therefore Exhibit C.21 also reports information for the complete case sample, or teachers and students with both baseline and outcome measures in the analytic sample. This information can be used to assess baseline equivalence. Exhibit C.22 shows the number of schools randomly assigned and numbers of students and teachers in these schools at key points in the study, to support calculations of cluster-level attrition and individual-level nonresponse. Finally, Exhibit C.23 presents the effects of study-provided coaching among the alternative sample of teachers present in study schools in the first six weeks of the school year. Results for this alternative sample are almost identical to the results for the main sample of teachers.

Exhibit C.21. Additional descriptive statistics for systematic reviews

	Control group				Five-cycle coaching group				Eight-cycle coaching group			
	Unadjusted mean	Unadjusted standard deviation	Number of students or teachers	Number of schools	Unadjusted mean	Unadjusted standard deviation	Number of students or teachers	Number of schools	Unadjusted mean	Unadjusted standard deviation	Number of students or teachers	Number of schools
Student achievement in math (z-scores)												
Outcome measure (analytic sample)	-0.15	0.95	3,058	37	-0.06	0.92	2,643	36	-0.12	0.95	2,475	34
Outcome measure (complete cases)	-0.14	0.95	2,847	37	-0.05	0.92	2,432	36	-0.11	0.95	2,280	34
Baseline measure (complete cases)	-0.13	0.92	2,847	37	-0.10	0.91	2,432	36	-0.07	0.95	2,280	34
Correlation between outcome and baseline measure	0.841											
Student achievement in English language arts (z-scores)												
Outcome measure (analytic sample)	-0.12	0.95	3,034	37	0.02	0.97	2,659	36	0.00	0.98	2,495	34
Outcome measure (complete cases)	-0.11	0.95	2,827	37	0.04	0.96	2,450	36	0.01	0.98	2,299	34
Baseline measure (complete cases)	-0.09	0.94	2,827	37	0.00	0.95	2,450	36	0.03	0.96	2,299	34
Correlation between outcome and baseline measure	0.806											
Teachers' overall CLASS score for main sample												
Outcome measure (analytic sample)	4.59	0.57	125	36	4.48	0.47	99	33	4.43	0.55	98	31
Outcome measure (complete cases)	4.60	0.57	123	36	4.48	0.47	99	33	4.43	0.55	96	31

	Control group				Five-cycle coaching group				Eight-cycle coaching group			
	Unadjusted mean	Unadjusted standard deviation	Number of students or teachers	Number of schools	Unadjusted mean	Unadjusted standard deviation	Number of students or teachers	Number of schools	Unadjusted mean	Unadjusted standard deviation	Number of students or teachers	Number of schools
Baseline measure (complete cases)	4.63	0.52	123	36	4.59	0.45	99	33	4.53	0.49	96	31
Correlation between outcome and baseline measure	0.268											36
Teachers' overall CLASS score for alternative sample												
Outcome measure (analytic sample)	4.60	0.57	123	36	4.48	0.47	99	33	4.43	0.55	97	31
Outcome measure (complete cases)	4.60	0.57	123	36	4.48	0.47	99	33	4.43	0.55	96	31
Baseline measure (complete cases)	4.63	0.52	123	36	4.59	0.45	99	33	4.53	0.49	96	31
Correlation between outcome and baseline measure	0.268											

Source: CLASS ratings of video-recorded classroom observations in spring 2019 and administrative student records for the 2017–2018 and 2018–2019 school years.

Note: The overall CLASS score and domain scores range from 1-7, with higher values indicating more positive outcomes. The overall CLASS score is an average of each of three domain scores– classroom management, building students' understanding, and building supportive relationships with students. Each domain score is the average of scores on a series of dimensions, which in turn are averages of specific indicator scores. The classroom management domain includes dimensions for behavior management, productivity, and negative climate. The building students' understanding domain includes dimensions for instructional learning formats, content understanding, analysis and inquiry, quality of feedback, and instructional dialogue. The building supportive relationships with students domain includes dimensions for positive climate, teacher sensitivity, and regard for adolescent perspectives. Because a confirmatory factor analysis indicated the data could not distinguish these last two domains, the study team combined them by averaging them. Test scores were converted to z-scores by subtracting the mean and dividing by the standard deviation of scores for all students in that state and grade level.

CLASS = Classroom Assessment Scoring System.

Exhibit C.22. Information needed to calculate attrition and nonresponse for systematic reviews

Count of individuals or clusters	Control group	Five-cycle coaching group	Eight-cycle coaching group
Number of schools randomly assigned	37	36	34
Number of math students in study schools at start of study school year	3,236	2,785	2,649
Number of English language arts students in study schools at start of study school year	3,184	2,793	2,641
Number of teachers in study schools six weeks into school year	132	116	110
Number of teachers in study schools at the time of end-of-year classroom observations	132	112	109

Source: Administrative student records for the 2017-2018 and 2018-2019 school years.

Exhibit C.23. Effects of the study-provided coaching on teachers' classroom practices, alternative sample

Outcome	Means			Effects					
				Five-cycle coaching group		Eight-cycle coaching group		Difference in effects of five and eight cycles	
	Control group	Five-cycle coaching group	Eight-cycle coaching group	Effect	*p*-value	Effect	*p*-value	Difference	*p*-value
Overall CLASS score	4.60	4.47	4.41	-0.14	0.11	-0.19*	0.03	0.050	0.51
Classroom management	6.46	6.28	6.19	-0.18*	0.01	-0.27*	0.00	0.090	0.19
Building students' engagement	5.38	5.25	5.21	-0.13	0.21	-0.17	0.11	0.040	0.62
Building students' understanding	3.57	3.43	3.40	-0.14	0.20	-0.17	0.11	0.030	0.78
Building supportive relationships with students	4.21	4.09	4.03	-0.11	0.36	-0.18	0.12	0.070	0.55
Number of teachers	**123**	**99**	**97**						
Number of schools	**36**	**33**	**31**						

Source: CLASS ratings of video-recorded classroom observations in spring 2019.

Note: The study estimated the effects for the five- and eight-cycle coaching groups by comparing outcomes for each of those groups to outcomes for the control group. The overall CLASS score and domain scores range from 1 to 7, with higher values indicating more positive outcomes. Each domain score is the average of scores on a series of dimensions. The classroom management domain includes dimensions for behavior management, productivity, and negative climate. The building students' understanding domain includes dimensions for instructional learning formats, content understanding, analysis and inquiry, quality of feedback, and instructional dialogue. The building supportive relationships with students domain includes dimensions for positive climate, teacher sensitivity, and regard for adolescent perspectives. The overall CLASS score is an average of 12 dimension-level scores—those that make up the three domains along with an additional stand-alone dimension for building student engagement, which also ranges from 1 to 7.

* Statistically significant at the .05 level, two-tailed test.

CLASS = Classroom Assessment Scoring System.

C.4 Minimum detectable effects

To summarize the level of precision in this study, Exhibit C.24 shows, for each key outcome, the realized values of the minimum detectable effects based on the study's actual data and approach. The minimum detectable effect is the smallest true effect for which the study had an 80 percent probability of obtaining an estimate that was statistically significant at the 5 percent level.

Exhibit C.24. Realized values of minimum detectable effects

Outcome	Estimated impact of five cycles	Minimum detectable impact of five cycles	Estimated impact of eight cycles	Minimum detectable impact of eight cycles
Student achievement (standard deviations)				
Student achievement in English language arts	0.08	0.08	0.02	0.09
Student achievement in math	0.07	0.10	0.02	0.11
Teachers' practices (points on scale from 1-7)				
Overall CLASS score	-0.13	0.24	-0.18	0.24
Classroom organization	-0.18	0.18	-0.26	0.19
Student engagement	-0.12	0.28	-0.15	0.29
Instructional support	-0.14	0.31	-0.17	0.30
Emotional support	-0.10	0.34	-0.17	0.32

Source: Administrative student records for the 2017-2018 and 2018-2019 school year, CLASS ratings of video-recorded classroom observations in spring 2019.

Note: The minimum detectable impact is the smallest true impact for which the study had an 80 percent probability of obtaining an estimate that was statistically significant at the 5 percent level. For each outcome, the study team calculated the minimum detectable impact by multiplying the standard error of the impact estimate by 2.8.

ENDNOTES

[1] In the CLASS rubric, classroom management practices comprise the Classroom Organization domain; practices related to building students' understanding comprise the Instructional Support domain; and practices related to building supportive relationships with students comprise the Emotional Support domain.

[2] These percentages are for the teachers who completed the intended number of cycles (five or eight), discussed further in Section A.2.2.

[3] Teachstone created the exemplar video clips. The clips show examples of effective practices for a specific CLASS dimension. Coaches assigned one clip for each CLASS dimension being addressed in the coaching cycle (for a total of two clips). The clips are short enough that the teacher can view them in 1 to 2 minutes.

[4] The study design was preregistered with the Registry of Efficacy and Effectiveness Studies, registry ID 1649.1, last updated on November 29, 2019.

[5] Hill et al. (2007).

[6] Pianta et al. (2012) provides evidence of the validity and reliability of the CLASS rubric from three studies, including the Measures of Effective Teaching project that includes a sample of 1,333 teachers across six districts (Kane and Staiger 2012).

[7] Kane and Staiger (2012) provides evidence of the validity and reliability of the PLATO rubric.

[8] Deng and Chan (2017).

[9] Hair et al. (2010).

[10] Kappa can report low reliability despite high agreement between raters in certain circumstances (Byrt, Bishop and Carlin 1993; Nurjannah and Siwi 2017; Zhao 2011). Gwet's AC1 statistic is designed to overcome the limitations of kappa (Xie 2013).

[11] Cohen (1960).

[12] Pianta et al. 2012.

[13] Graham et al. 2012.

[14] The study's design registry (registry ID 1649.1) designated six confirmatory analyses—the effects of five cycles of coaching on students' math scores, students' English language arts scores, and teachers' general classroom practices (measured by overall score on the CLASS), and the effects of eight cycles of coaching on these same outcomes. Because the study team considered each of these outcomes to be separate domains, the team did not adjust for multiple hypothesis tests across these six analyses, following the guidance from Schochet (2008). All other analyses looking at effects on other outcomes or for particular subgroups are considered exploratory. Therefore, the team also did not adjust for multiple hypothesis testing in these exploratory analyses, following the guidance from Schochet (2008).

[15] Puma et al. (2009).

[16] Hollands et al. (2015).

[17] The cost effectiveness of the strategies shown in Exhibit B.20 is based on studies of the effect of teacher pay-for-performance as implemented by the U.S. Department of Education's Teacher Incentive Fund grantees (Chiang et al. 2017); a reduction in class size from classes with 22 to 26 students to classes with 13 to 17 students (Nye et al. 2000); and $20,000 over two years for high-performing teachers who transferred into low-performing schools (Glazerman et al. 2013).

[18] Estimated effects of the coaching are similar in sign and magnitude to those in Exhibit C.1 if the analysis adjusts only for randomization-block fixed effects and no other observable characteristics.

[19] Estimated effects of the coaching are similar in sign and magnitude to those in Exhibit C.8 if the analysis adjusts only for randomization-block fixed effects and no other observable characteristics.

REFERENCES

Byrt, Ted, Janet Bishop, and John Carlin. "Bias, Prevalence and Kappa." *Journal of Clinical Epidemiology*, vol. 46, no. 5, 1993, pp. 423–429.

Chiang, Hanley, Cecilia Speroni, Mariesa Herrmann, Kristin Hallgren, Paul Burkander, and Alison Wellington. "Evaluation of the Teacher Incentive Fund: Final Report on Implementation and Impacts of Pay for Performance Across Four Years (NCEE 2017–4004)." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2017.

Cohen, Jacob. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement*, vol. 20, no. 1, 1960, pp. 37–46.

Deng, Lifang, and Wai Chan. "Testing the Difference Between Reliability Coefficients Alpha and Omega." *Educational and Psychological Measurement*, vol. 77, no. 2, 2017, pp. 185–203.

Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, and Jeffrey Max. "Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment." NCEE 2013–4003. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, November 2013.

Graham, Matthew, Anthony Milanowski, and Jackson Miller. "Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings." Madison, WI: Center for Educator Compensation Reform, February 2012.

Gwet, Kilem Li. *Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Gaithersburg, MD: Advanced Analytics, LLC, 2012.

Hair, Joseph, William Black, Barry Babin, and Rolph Anderson. *Multivariate Data Analysis*. London, United Kingdom: Pearson College Division, 2010.

Hill, Carolyn, Howard Bloom, Alison Rebeck Black, and Mark Lipsey. "Empirical Benchmarks for Interpreting Effect Sizes in Research." New York, NY: MDRC, July 2007.

Hollands, Fiona, Barbara Hanisch-Cerda, Henry Levin, Clive Belfield, Amritha Menon, Robert Shand, Yilin Pan, Ipek Bakir, and Henan Cheng. CostOut - the CBCSE Cost Tool Kit. Center for Benefit-Cost Studies of Education, Teachers College, Columbia University, 2015. Available at www.cbcsecosttoolkit.org. Accessed October 19, 2020.

Kane, Thomas, and Douglas Staiger. "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains." Seattle, WA: Bill & Melinda Gates Foundation, January 2012.

Nurjannah, Intansari, and Sri Marga Siwi. "Guidelines for Analysis on Measuring Interrater Reliability of Nursing Outcome Classification." *International Journal of Research in Medical Sciences*, vol. 5, no. 4, 2017, pp. 1169–1175.

Nye, Barbara, Larry V. Hedges, Spyros Konstantopoulos. "The Effects of Small Classes on Academic Achievement: The Results of the Tennessee Class Size Experiment." *American Educational Research Journal*, Spring 2000, vol. 37, No. 1, pp. 123–151.

Pianta, Robert, Bridget Hamre, and Susan Mintz. "Upper Elementary and Secondary CLASS Technical Manual." Charlottesville, VA: University of Virginia, 2012.

Puma, Michael J., Robert B. Olsen, Stephen H. Bell, and Cristofer Price. "What To Do When Data Are Missing in Group Randomized Controlled Trials." NCEE 2009-0049. Washington, DC: National Center for Education Evaluation and Regional Assistance, 2009.

Schochet, Peter Z. "Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations." NCEE 2008-4018. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2008.

What Works Clearinghouse. *What Works Clearinghouse Standards Handbook, Version 4.1.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, 2020. Available at https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf. Accessed March 3, 2022.

Xie, Quingshu. "Agree or Disagree? A Demonstration of An Alternative Statistics Cohen's Kappa for Measuring the Extent and Reliability of Agreement Between Observers." Unpublished manuscript, 2013. Available at https://nces.ed.gov/FCSM/pdf/J4_Xie_2013FCSM.pdf. Accessed May 21, 2021.

Zhao, Xinshu. "When to Use Cohen's K, If Ever?" International Communication Association Conference, Boston, Massachusetts, May 2011.