Making Connections

November 2016

The content, predictive power, and potential bias in five widely used teacher observation instruments

Brian Gill
Megan Shoji
Thomas Coen
Kate Place
Mathematica Policy Research

Key findings

This study seeks to inform decisions about the selection and use of teacher observation instruments using data from the Measures of Effective Teaching project. It compares five widely used observation instruments on the practices they measure, their relationship to student learning, and whether they are affected by the characteristics of students in a teacher's classroom. The study found that:

- Eight of ten dimensions of instructional practice are common across all five examined teacher observation instruments.
- All seven of the dimensions of instructional practice with quantitative data are modestly but significantly related to teachers' value-added scores.
- The classroom management dimension is most consistently and strongly related to teachers' value-added scores across instruments, subjects, and grades.
- The characteristics of students in the classroom affect teacher observation results for some instruments, more often in English language arts classes than in math classes.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

REL MID-ATLANTIC
Regional Educational Laboratory
At ICF International

U.S. Department of Education
John B. King, Jr., *Secretary*

Institute of Education Sciences
Ruth Neild, *Deputy Director for Policy and Research*
Delegated Duties of the Director

National Center for Education Evaluation and Regional Assistance
Joy Lesnick, *Acting Commissioner*
Amy Johnson, *Action Editor*
Felicia Sanders, *Project Officer*

REL 2017–191

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

November 2016

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0006 by Regional Educational Laboratory Mid-Atlantic administered by ICF International. 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.

Gill, B., Shoji, M., Coen, T., and Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments* (REL 2017–191). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs.

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Summary

School districts and states across the Regional Educational Laboratory Mid-Atlantic Region and the country as a whole have been modifying their teacher evaluation systems to identify more effective and less effective teachers and provide better feedback to improve instructional practice. The new systems typically include components related to student achievement growth and instruments for observing and rating instructional practice.

Many school districts and states are considering adopting commercially available instruments for the instructional practice component of their evaluation systems. Yet little data are available to help districts and states choose among available instruments or determine which dimensions of instructional practice merit the greatest emphasis. Most existing data comparing different observation instruments, including their statistical characteristics and their relationship to student achievement, come from the Bill & Melinda Gates Foundation's Measures of Effective Teaching project (Kane & Staiger, 2012).

This study examined data from the Measures of Effective Teaching project to address three research questions that might inform district and state decisions about selecting and using five widely used teacher observation instruments: the Classroom Assessment Scoring System, the Framework for Teaching, the Protocol for Language Arts Teaching Observations, the Mathematical Quality of Instruction, and the UTeach Observational Protocol. Specifically, the research questions focused on the major differences and similarities in the dimensions of instructional practice rated by the five observation instruments, whether some dimensions of instructional practice consistently show stronger correlations with teachers' value-added scores across the different observation instruments, and the extent to which characteristics of students in the classroom affect instrument scores.

Key findings include:

- Eight of ten dimensions of instructional practice are common across all five examined teacher observation instruments, demonstrating that large parts of the various instruments are conceptually consistent.
- All seven of the dimensions of instructional practice with quantitative data are modestly but significantly related to teachers' value-added scores.
- The classroom management dimension is most consistently and strongly related to teachers' value-added scores across instruments, subjects, and grades.
- The characteristics of students in the classroom affect teacher observation scores for some instruments and subjects. Observation scores for English language arts classes may be more susceptible to classroom composition effects. For two of the three instruments (Framework for Teaching and Classroom Assessment Scoring System) used to score English language arts instruction, teachers with a larger percentage of racial/ethnic minority students in their classroom tend to receive lower observation scores; a similar effect was observed with the Framework for Teaching for teachers with lower-achieving students. There was no evidence that the composition of students in the classroom affects scores for the Protocol for Language Arts Teaching Observations (the third instrument used to score English language arts instruction), and there was little indication that student characteristics affect observation scores in math classes.

Contents

Why this study?

School districts and states across the Regional Educational Laboratory (REL) Mid-Atlantic Region and the country as a whole have been modifying their teacher evaluation systems, with considerable federal support. The new systems aim to better identify more effective and less effective teachers and provide better feedback to improve instructional practice.

Student achievement growth is a major component of new evaluation systems. Many districts and states are using statistical methods designed to measure teachers' contributions to student achievement growth, known as value-added models (which calculate teachers' impacts on student achievement after accounting for the influence of other student factors, including prior achievement and demographic characteristics) or student growth percentiles (which express student achievement gains relative to students who started out at a similar level). But value-added models and student growth percentiles are not a panacea for teacher evaluation. U.S. Department of Education and state policies indicate that value-added models are neither a complete nor comprehensive measure of teacher effectiveness (Doherty & Jacobs, 2013). Consequently, schools, districts, and states across the country are seeking to improve other components of the teacher evaluation system—in particular, observation-based measures of teachers' instructional practice.

Many districts and states are considering commercially available observation instruments for the instructional practice component of their systems. Yet little data are available to help districts and states choose an instrument or identify particular dimensions of instructional practice

Many districts and states are considering commercially available observation instruments for the instructional practice component of their systems. Yet little data are available to help districts and states choose an instrument or identify particular dimensions of instructional practice that may merit the most weight in evaluation and the most attention in professional development. Most of the research comparing observation instruments, including their statistical characteristics and their relationship to student achievement, is from the Bill & Melinda Gates Foundation's Measures of Effective Teaching project (Kane & Staiger, 2012). In response to a request by the REL Mid-Atlantic Teacher Evaluation Research Alliance, the current study examined data from the Measures of Effective Teaching project on five observation instruments used to measure teacher instructional practice (table 1).

Research has found that overall scores on observation instruments predict student learning as measured by test scores. Prior analyses of Measures of Effective Teaching project data found that overall scores from the five observation instruments are correlated with teachers' value-added scores (Kane & Staiger, 2012; see also Garrett & Steinberg, 2015)

Table 1. Teacher observation instruments used in this study

Observation instrument	Developed by	Type of classes served
Classroom Assessment Scoring System	University of Virginia	English language arts and math
Framework for Teaching	Charlotte Danielson	English language arts and math
Protocol for Language Arts Teaching Observations	Stanford University	English language arts
Mathematical Quality of Instruction	University of Michigan	Math
UTeach Observational Protocol	University of Texas–Austin	Math

Note: The Measures of Effective Teaching project also included the Quality Science Teaching instrument, which was excluded here because it is designed only for science classes and the current study focused only on English language arts and math instruction.

Source: Kane & Staiger, 2012.

and that the relationship between observation scores and teachers' value-added scores varies across school districts (Lynch, Chin, & Blazar, 2013). Similar evidence of significant correlations with students' test scores or with teachers' value-added scores has been documented in Pittsburgh, Pennsylvania, using an observation measure of instructional practice derived from the Framework for Teaching instrument (Chaplin, Gill, Thompkins, & Miller, 2014) and in New York City, using the Protocol for Language Arts Teaching Observations instrument and parts of the Classroom Assessment Scoring System instrument (Grossman, Loeb, Cohen, & Wyckoff, 2013). The latter study also concluded that focusing on improving particular observable teaching dimensions could lead to student achievement gains (Grossman et al., 2013).

While these studies suggest that composite measures for some existing observation instruments predict student achievement, they have not fully examined how relationships with student learning vary for different dimensions of instructional practice, such as classroom management versus content understanding. Such information might be useful to schools and districts that prioritize particular dimensions of instructional practice over others. This study provides new insights by summarizing the different dimensions of instructional practice measured in five teacher observation instruments and by exploring how observation scores in these dimensions are associated with teacher contributions to student learning as measured by teachers' value-added scores. One other recent study has conducted similar analyses using different data (Lockwood, Savitsky, & McCaffrey, 2015).

This study summarizes the different dimensions of instructional practice measured in five teacher observation instruments and explores how observation scores in these dimensions are associated with teacher contributions to student learning as measured by teachers' value-added scores

Because teachers work with students of diverse backgrounds and abilities, decisions about observation instruments and their instructional dimensions could also be informed by whether instrument scores are affected by the students in the classroom. Such a relationship would emerge if teachers ineffectively alter their instruction depending on the characteristics of students they teach. For example, one study found that White teachers have significantly lower educational expectations of Black students in grade 10 than Black teachers do for the same students (Gershenson, Holt, & Papageorge, 2015). Classroom composition would also influence instrument scores if instruments are implicitly biased against teachers serving students with particular characteristics. This would imply that instruments partially rate teachers on the characteristics of students they teach rather than solely on their instructional practices.

The literature on the relationship between observation scores and student characteristics is small. A substantial methodological literature assesses whether and how value-added models can fully account for the characteristics of students served by different teachers, and there is growing evidence that value-added models can largely succeed in doing so (Chetty, Friedman, & Rockoff, 2014; Goldhaber & Chaplin, 2015; Kane, McCaffrey, Miller, & Staiger, 2013). Observation measures of teachers' instructional practice have not been subject to the same scrutiny. Few studies have examined the relationship between observation measures of instructional practice and student characteristics. Several studies have found that teachers of racial/ethnic minority, lower-income, or lower-achieving students tend to have lower observation scores, but the studies rely primarily or exclusively on data that cannot determine whether the finding is an effect of the instrument or a result of students with certain characteristics being assigned to teachers who are less effective (Borman & Kimball, 2005; Chaplin et al., 2014; Whitehurst, Chingos, & Lindquist, 2014). This study sheds light on the source of the relationship between observation scores and student characteristics by leveraging random assignment of teachers to classrooms in the

Measures of Effective Teaching project, allowing the study team to assess the causal effect of student characteristics on teacher observation scores.[1]

This study may help school district and state decisionmakers weigh the benefits and drawbacks of five widely used observation instruments by considering which dimensions of instructional practice each instrument measures, how different dimensions of instructional practice are related to student learning, and whether observation measures are related to the characteristics of students in a teacher's classroom.

What the study examined

This study was guided by three research questions:

- What are the major differences and similarities in the dimensions of instructional practice rated by five teacher observation instruments?
- Across observation instruments, do some dimensions of instructional practice consistently show stronger correlations with teachers' value-added scores?
- To what extent do the characteristics of students in the classroom affect observation scores? Are ratings on some instruments, or for some dimensions of instructional practice, more influenced by student characteristics than others?

This study may help school district and state decisionmakers weigh the benefits and drawbacks of five widely used observation instruments by considering which dimensions of instructional practice each instrument measures, how different dimensions of instructional practice are related to student learning, and whether observation measures are related to the characteristics of students in a teacher's classroom

To answer these questions, the study examined the content and statistical strengths and weaknesses of five observation instruments used in the Measures of Effective Teaching project (see table 1).

Collectively, the examinations provide comparative information that districts and states can use to select appropriate instruments for their teacher evaluation systems and to identify dimensions of instructional practice that might merit particular emphasis. See box 1 for a brief description of the data and methods and appendix A for more detail.

Box 1. Data and methods

Instrument content analysis of dimensions of instructional practice

The study team first reviewed the rubrics of five observation instruments—the Classroom Assessment Scoring System, the Framework for Teaching, the Protocol for Language Arts Teaching Observations, the Mathematical Quality of Instruction, and the UTeach Observational Protocol—to identify the specific instructional practices measured by each instrument. The study team used qualitative coding (content analysis) to classify text on specfic practices across all five instruments into common dimensions of instructional practice.

Correlation analysis of teachers' observation scores and value-added scores

The study team then estimated correlations between teachers' scores for each observation-instrument dimension and teachers' value-added scores, using data on grade 4–9 English language arts and math teachers from the Measures of Effective Teaching project's longitudinal database. The database, housed at the Inter-University Consortium for Political and Social Research, was compiled by the University of Michigan and contains teacher quality data on more than 2,500 grade 4–9 teachers who volunteered to participate for the 2009/10 and

(continued)

Box 1. Data and methods *(continued)*

2010/11 school years. Participating teachers represented 317 schools across six school districts: Charlotte-Mecklenburg County, North Carolina; Dallas, Texas; Denver, Colorado; Hillsborough County, Florida; Memphis, Tennessee; and New York, New York. The data contain student-level and classroom-level test scores and demographic information, as well as teachers' observation scores on the five instruments.

The study team calculated teachers' value-added scores using value-added models that estimated teacher contributions to student achievement gains on state-administered standardized tests in English language arts and math. Using teachers' observation subscores for the five instruments, the study team constructed observation dimension scores in seven dimensions of instructional practice.

The study team calculated both instrument-specific dimension scores and a cross-instrument average dimension score for each dimension. Correlations were estimated, separately by subject (English language arts or math) and by grade-level groupings (grades 4–5 versus grades 6–9), between teachers' value-added scores and teachers' instrument-specific and cross-instrument observation dimension scores.

Correlation analysis of teacher observation scores and student characteristics

Finally, the study team examined the extent to which instruments' total ratings and dimension scores are influenced by the characteristics of students in the classroom. This required more than a simple correlation of teacher observation scores with student characteristics, because teachers are not (usually) randomly assigned to classrooms. Simple correlations could be biased if less effective teachers were more likely to be assigned to classrooms with a larger percentage of lower-achieving, lower-income, or racial/ethnic minority students, as a result of seniority privileges or other factors affecting teacher assignments. To bypass this problem, the study analyzed a subset of teachers in the Measures of Effective Teaching project who were randomly assigned to student classroom groups within their schools in the second year of the Measures of Effective Teaching project.

Within the randomized sample of teachers, the study team conducted a series of regression analyses, examining the relationship between teachers' instrument-specific observation scores and the characteristics of students who were randomly assigned to each classroom and remained in the Measures of Effective Teaching project. Four measures of classroom-level student characteristics were used:

- Percentage of students who are eligible for the federal school lunch program (a proxy for low-income status).
- Percentage of students who are racial/ethnic minority students.
- Average prior-year English language arts test score (on state-administered standardized tests).
- Average prior-year math test score (on state-administered standardized tests).

The study estimated two sets of regressions using different types of observation scores: instrument-specific overall scores (total score for each instrument, standardized) and instrument-specific dimension scores in seven dimensions of instructional practice (see above).

What the study found

This section describes findings from the analyses of instrument content and teacher observation scores from the Measures of Effective Teaching project. The study identified 10 key dimensions of instructional practice, 8 of which were rated by all five instruments. All seven dimensions of instructional practice with quantitative data were modestly related to teachers' value-added scores. Classroom management showed the strongest and most consistent correlations to teachers' value-added scores across instruments, subjects, and grades. The characteristics of students in the classroom affected teacher observation results for some instruments, more often in English language arts classes than in math classes.

Instrument content analysis

The study identified 10 key dimensions of instructional practice, 8 of which were rated by all five instruments. Classroom management showed the strongest and most consistent correlations to teachers' value-added to student learning across instruments, subjects, and grades

To explore the similarities and differences in the dimensions of instructional practice measured by different observation instruments, the study team conducted a content analysis of the rubrics of the Classroom Assessment Scoring System, Framework for Teaching, Protocol for Language Arts Teaching Observations, Mathematical Quality of Instruction, and UTeach Observational Protocol. The analysis yielded 10 key dimensions of instructional practice that are rated by at least one instrument, and 8 of those dimensions are rated by all five instruments (box 2; table 2).

Box 2. Ten dimensions of instructional practice rated in the five observation instruments

The analysis yielded evidence of 10 key dimensions of instructional practice. Eight of these dimensions are rated in all five instruments; they are identified with an asterisk.

- *Supportive learning environment.** Degree to which the teacher and students display warmth, enjoyment, praise, and respect in their interactions; the classroom is free of negativity (as created through, for example, yelling, bullying, physical aggression, or disrespectful language from the teacher or students); the teacher demonstrates awareness and responsiveness to student needs; the environment is inclusive of all students regardless of background or ability; students demonstrate comfort in sharing ideas, asking questions, or otherwise participating; and the teacher imparts high expectations for student work and establishes norms or guidelines for supportive feedback and student interactions.
- *Student focus.** Degree to which the teacher encourages student ideas, actively listens to and comments on student responses, incorporates student ideas into the lesson, demonstrates flexibility, or adjusts the lesson to students' understanding or ability; the students have responsibilities, choices, and leadership opportunities; each student has a role and participates in group work; and, when appropriate, the teacher engages families to support student development.
- *Classroom management.** Degree to which behavioral rules and expectations are clear; the teacher anticipates or effectively redirects misbehavior; students behave well; the teacher manages time effectively (for example, by minimizing disruptions, providing tasks for students, minimizing time spent on administrative tasks, or preparing the lesson and materials in advance); and students manage time effectively (for example, by knowing what they have to do, efficiently transitioning from one activity to another, or being on task).

(continued)

Box 2. Ten dimensions of instructional practice rated in the five observation instruments *(continued)*

- *Active student participation in class activities.* Degree to which students are actively participating in class activities most of the time.
- *Student intellectual engagement with content.** Degree to which students intellectually engage with material; tasks require higher-order thinking skills and are sufficiently challenging to push students cognitively; students engage in open-ended tasks, analysis, prediction, or interpretation; and students ask questions that demonstrate thought, provide alternative strategies or challenge statements, or explain an approach, the meaning of an answer, or a thought process.
- *Lesson structure and facilitation.** Degree to which the teacher communicates learning objectives and presents the lesson in a clear and well-organized way, using a variety of strategies or types of materials and appropriate pacing to allow time for summary or reflection at the end; the teacher uses active facilitation techniques to engage all students; and the teacher arranges the room in a way that supports learning and safety.
- *Content understanding.** Degree to which the lesson content is meaningful and important to learn; the teacher demonstrates deep content understanding; the presentation of content and written materials do not contain errors; the teacher provides explicit explanation of a procedure or skill and provides students the opportunity to practice applying it; the teacher provides in-depth explanations, uses multiple examples, and spends time responding to student questions to improve understanding of the content; and the teacher connects the content to other disciplines or examples, the real world, or students' prior knowledge or personal experiences.
- *Language and discourse.** Degree to which the teacher and students facilitate discussion through open-ended questions, active listening, acknowledgement, pauses, and the like; students take a lead role in discussions; conversations build on each other, with students responding to one another; and academic or technical vocabulary is defined, encouraged, and used often.
- *Feedback and assessment.** Degree to which feedback is specific, in-depth, and helps advance student understanding; feedback includes back-and-forth exchanges between teacher and student, scaffolding or hints to students, or correction of student misconceptions; the teacher creates opportunities to formally or informally assess student understanding of content; and formative and summative assessments are linked to learning objectives.
- *Teacher professionalism.* Degree to which the teacher is reflective of his or her instructional practice; exhibits an ability to identify strengths and weaknesses; actively engages in learning opportunities to improve teaching skills; collaborates with or demonstrates mutual trust and rapport with parents and colleagues; and participates in community activities or school decisionmaking.

Table 2. Dimensions of instructional practice rated by each teacher observation instrument

Dimension	Classroom Assessment Scoring System	Framework for Teaching	Protocol for Language Arts Teaching Observations	Mathematical Quality of Instruction	UTeach Observational Protocol
Supportive learning environment	✔	✔	✔	✔	✔
Student focus	✔	✔	✔	✔	✔
Classroom management	✔	✔	✔	✔	✔
Active student participation in class activities	✔		✔	✔	
Student intellectual engagement with content	✔	✔	✔	✔	✔
Lesson structure and facilitation	✔	✔	✔	✔	✔
Content understanding	✔	✔	✔	✔	✔
Language and discourse	✔	✔	✔	✔	✔
Feedback and assessment	✔	✔	✔	✔	✔
Teacher professionalism		✔			✔

Note: See table C3 in appendix C for example instrument text coded for each dimension.

Source: Authors' content analysis of the rubrics of the five teacher observation instruments.

To select the observation instrument that best meets their needs, districts and states might consider which instruments measure particular dimensions of instructional practice they wish to target

The five instruments differ in the subdimensions of each dimension that they measure (see table C1 in appendix C for a list of the subdimensions covered in each instrument and table C2 for definitions of each subdimension). For example, within the supportive learning environment dimension, most instruments rate how aware and responsive teachers are to student needs (teacher awareness and responsiveness) and how comfortable students appear (student ease in educational environment), while only two rate inclusiveness of all students regardless of background or ability (inclusive class environment) or whether the teacher holds students to high standards (high expectations for students).

To select the observation instrument that best meets their needs, districts and states might consider which instruments measure particular dimensions of instructional practice they wish to target. For example, the Framework for Teaching instrument might be preferred by districts and states that prioritize how well schools facilitate parental involvement, as it is the only instrument with content capturing family engagement (within the student focus dimension). In contrast the Classroom Assessment Scoring System or Protocol for Language Arts Teaching Observations instruments might be preferred by districts and states that believe that educators' ability to facilitate discussion is crucial for students to develop critical thinking skills, as these are the only instruments with content capturing discussion facilitation and cumulative exchanges (within the language and discourse dimension).

Instruments also differ in how comprehensively they measure instructional practice in each dimension. On average the instruments cover 57–75 percent of the subdimensions in a given dimension of instruction (table 3). The Protocol for Language Arts Teaching Observations tends to exhibit the best coverage of subdimensions, rating the highest percentage of subdimensions in 7 of the 10 dimensions.

Table 3. Percentage of subdimensions rated in each dimension of instruction, by observation instrument

Dimension (number of subdimensions)	Classroom Assessment Scoring System	Framework for Teaching	Protocol for Language Arts Teaching Observations	Mathematical Quality of Instruction	UTeach Observational Protocol
Supportive learning environment (5)	60	80	80	20	80
Student focus (4)	75	75	75	50	50
Classroom management (2)	100	100	100	100	100
Active student participation in class activities (1)	100	0	100	100	0
Student intellectual engagement with content (4)	25	100	75	75	25
Lesson structure and facilitation (4)	75	100	50	75	100
Content understanding (6)	83	50	100	100	83
Language and discourse (3)	67	33	100	67	33
Feedback and assessment (6)	67	50	67	50	67
Teacher professionalism (3)	0	100	0	0	33
Mean across all dimensions	65	69	75	64	57

Note: See table C1 in appendix C for list of subdimensions covered in each instrument and table C2 for definitions of each subdimension.

Source: Authors' content analysis of observation instrument rubrics.

Information on observation instruments could be used to assess which instruments maximize coverage of subdimensions within the instructional dimensions of particular interest to districts or states that are selecting an observation instrument

This information could be used to assess which instruments maximize coverage of subdimensions within the instructional dimensions of particular interest to districts or states that are selecting an observation instrument. For example, districts or states concerned with comprehensively assessing teacher performance within the student intellectual engagement with content and teacher professionalism dimensions might prefer the Framework for Teaching instrument. However, that instrument would be less ideal for districts and states prioritizing comprehensive assessment of teachers' content understanding and language and discourse, as it covers fewer subdimensions than others within those dimensions; the Protocol for Language Arts Teaching Observations instrument would provide the richest measures for that set of priorities.

Relationships between teacher observation scores and value-added scores

To explore the extent to which observation scores in some dimensions of instructional practice consistently show stronger correlations with teachers' value-added scores, the study team estimated correlations between teachers' observation dimension scores (constructed from subscores defined by instrument developers) and teachers' contributions to student learning on state assessments in English language arts and math (estimated in value-added models).

To control for the idiosyncrasies of particular instruments, the analysis used observation data only on the dimensions of instructional practice for which the Measures of Effective Teaching longitudinal database included pre-existing subscores from more than one instrument. Although all 10 dimensions of instructional practice identified in the content analysis were represented by at least one subscore in the Measures of Effective Teaching

longitudinal database, only seven were measured by pre-existing subscores from more than one instrument. Consequently, the study team investigated the relationships between teachers' value-added scores and teachers' scores on these seven dimensions of instructional practice: supportive learning environment, classroom management, student intellectual engagement with content, lesson structure and facilitation, content understanding, language and discourse, and feedback and assessment.

To assess the strength of associations with teachers' value-added scores for different dimensions of instructional practice, the study team correlated teachers' value-added scores and their cross-instrument average dimension scores. The overall dimension score is the mean of the standardized instrument-specific dimension scores for each eligible instrument within the dimension group (see table A6 in appendix A). The pattern of results is the same when averages of instrument-specific dimension scores are compared. The study team then used the year-to-year variation in teachers' value-added scores to produce an adjusted correlation that may be interpreted as the correlation between teachers' average observation dimension score and their underlying value added—the value added that is stable for a teacher over time, rather than a single-year measure (Kane & Staiger, 2012; see appendix A for details).

Average teacher observation scores on the seven dimensions of instructional practice for which sufficient data were available showed statistically significant relationships with teachers' value-added scores

Average teacher observation scores on the seven dimensions of instructional practice for which sufficient data were available showed statistically significant relationships with teachers' value-added scores (with both sets of scores pooled across grade levels; based on a two-tailed significance test at the .05 level). The results indicate modest relationships between teachers' cross-instrument dimension scores and teachers' underlying value-added scores (table 4). Estimated correlations were largest for the classroom management dimension (with an adjusted correlation to underlying value added of .28). Results for particular grades and subjects are reported in table C5 in appendix C.

To assess consistency of associations with teachers' value-added scores for different instruments within a dimension, the study team estimated correlations between teachers' value-added scores and their instrument-specific dimension scores, separately by grade-level groups (grades 4–5 or grades 6–9) and subject (English language arts or math), and recorded the percentage of correlation estimates in each dimension that are statistically distinguishable from zero, using a significance level of .05 (table 5).

All seven examined dimensions had more statistically significant relationships to teachers' value-added scores than would be expected by chance (see table 5). Instrument-specific dimension scores were most often related to teachers' value-added scores for the classroom management and feedback and assessment dimensions of instructional practice.

To allow for similar comparisons across instruments, summary results were examined across all available dimension scores for each instrument (see table C6 in appendix C). All five instruments had more statistically significant relationships to value-added scores than would be expected by chance, with 17–47 percent of instruments' observation dimension scores significantly correlated with value-added scores. Among the five instruments, observation scores for the UTeach Observational Protocol instrument were most strongly and significantly correlated with teachers' value-added scores. These comparisons should be interpreted with caution as there were different numbers of correlation coefficients available for different instruments—depending on the number of dimension scores and years

Table 4. Summary results for the strength of relationship between teachers' cross-instrument observation dimension scores and their value added to student learning

Dimension	Adjusted correlation to underlying value added score
Supportive learning environment	.18
Classroom management	.28
Student intellectual engagement with content	.22
Lesson structure and facilitation	.18
Content understanding	.13
Language and discourse	.14
Feedback and assessment	.20

Note: Sample includes all teachers who taught a class with at least one valid observation instrument score and who taught at least five students with a valid state assessment outcome score. Grade 4–5 teachers in district 4 were excluded from these analyses because data on eligibility for the federal school lunch program were missing for all students. Results are reported for only 7 of the 10 dimensions identified in the content analysis because the scores for the other 3 dimensions were available for only one instrument. Reported results summarize correlations between teachers' value-added scores and their cross-instrument dimension score, adjusted for measurement error (see appendix A for details). All correlations shown were statistically distinguishable from zero (two-tailed test at the .05 level). The analysis estimated correlations separately by subject (English language arts or math) and primary and secondary grade levels (grades 4–5 or grades 6–9). For grades 4–5 correlations were estimated between value-added scores in one year and observation scores in another year for all available year combinations (year 1 value-added scores with year 2 observation scores, and vice versa). The reported results summarize findings across all grades and subjects as the mean of the adjusted rho estimates for each of the four subject-by-grade combinations: grade 4–5 English language arts, grade 4–5 math, grade 6–9 English language arts, and grade 6–9 math, weighted by the number of teachers included in each subject-by-grade group. See table C5 in appendix C for supplementary results, presented by subject and grade level.

Source: Authors' analysis of Measures of Effective Teaching data.

Table 5. Summary results for the consistency of the relationship between teachers' instrument-specific dimension scores and their value added to student learning for all subjects and grade levels

Dimension	Total number of correlations	Number of correlations that are significant	Percentage of correlations that are significant
Supportive learning environment	14	4	29
Classroom management	17	10	59
Student intellectual engagement with content	14	4	29
Lesson structure and facilitation	20	6	30
Content understanding	14	4	29
Language and discourse	17	4	24
Feedback and assessment	17	10	59

Note: Sample includes all teachers who taught a class with at least one valid observation instrument score and who taught at least five students with a valid state assessment outcome score. Grade 4–5 teachers in district 4 were excluded from these analyses because data on eligibility for the federal school lunch program were missing for all students. Significance is based on a two-tailed test at the .05 level.

Source: Authors' analysis of Measures of Effective Teaching data.

of data available for each—and it is possible that summary information based on fewer available correlations is less precise and more sensitive to outlier estimates.

In sum, while observation scores predict teachers' value-added scores for all seven dimensions of instructional practice examined and for all five instruments considered, scores in the classroom management dimension and for the UTeach Observational Protocol instrument stand out with the strongest and most consistent relationships.

Relationships between teacher observation scores and student characteristics

To explore the extent to which teacher observation scores differ systematically depending on the composition of students in a classroom, the study team estimated associations between observation scores[2] and average student characteristics of the assigned classroom by estimating regression coefficients. Using a subset of the randomized sample as described in appendix A, the study team examined relationships between observation instrument scores (overall and by dimension) and four student composition measures of assigned classrooms: percentage of students who are racial/ethnic minority students, percentage of students who are eligible for the federal school lunch program, prior-year average English language arts scores, and prior-year average math scores. In practice the actual students who ended up in a classroom did not always correspond perfectly with those randomly assigned (that is, there was some noncompliance with random assignment). The analyses use the characteristics of the randomly assigned classrooms rather than the actual classrooms to maintain the integrity of the experiment (producing intent-to-treat results), but the deviations from random assignment mean that the analyses underestimate the full effects of student characteristics on observation scores.

Examination of the relationships between observation instrument scores (overall and by dimension) and four student composition measures of assigned classrooms found 18 statistically significant results out of 156 regression combinations, more than double the number that would be expected by chance

For English language arts classrooms the study team investigated the relationship between each classroom characteristic and observation scores on the subject-specific Protocol for Language Arts Teaching Observations instrument, and the two subject-general instruments, the Classroom Assessment Scoring System and the Framework for Teaching. Similarly, they investigated relationships for math classrooms using scores on the subject-specific Mathematical Quality of Instruction instrument, as well as the Classroom Assessment Scoring System and Framework for Teaching instruments. In the study sample all 457 English language arts classrooms were scored on all three English language arts instruments and all 396 math classrooms were scored on all three math instruments.[3] The UTeach Observational Protocol instrument was not included in this analysis because it was not administered in randomized classrooms. Results are averaged across grade-level groups (grades 4–5 and grades 6–9). Thus, the study team estimated regression models predicting teachers' overall or dimension scores on an observation instrument as a function of a classroom composition characteristic, controlling for randomization block. Observation scores are standardized by subject.

The analyses found 18 statistically significant results out of 156 regression combinations (12 percent)—more than double the number that would be expected by chance.[4] These significant findings were clustered among three types of analyses (table 6; see table C7 for complete results). First, significant effects were prevalent beyond chance expectations in analyses of English language arts classrooms (19 percent of 80 total regressions) but not math classrooms (4 percent of 76 total regressions). Second, significant effects were more common for the percentage of students who are racial/ethnic minority students in a

Table 6. Statistically significant results for the relationship between teacher observation scores and classroom composition

Student characteristic, subject, and instrument	Dimension	Point estimate	Standard error	*p* value	Sample size
Average baseline English language arts score					
English language arts					
Framework for Teaching	Overall	0.28**	0.11	<.01	457
	Supportive learning environment	0.23*	0.10	.03	457
	Feedback and assessment	0.21*	0.09	.02	457
	Lesson structure and facilitation	0.37**	0.11	<.01	457
	Language and discourse	0.31**	0.12	.01	457
Math					
Mathematical Quality of Instruction	Lesson structure and facilitation	–0.33*	0.14	.02	355
Percentage of students who are racial/ethnic minority students					
English language arts					
Classroom Assessment Scoring System	Overall	–0.50*	0.24	.04	457
	Classroom management	–0.38*	0.19	.04	457
	Supportive learning environment	–0.59*	0.24	.02	457
Framework for Teaching	Overall	–0.71**	0.23	<.01	457
	Supportive learning environment	–0.74**	0.22	<.01	457
	Classroom management	–0.46*	0.20	.02	457
	Feedback and assessment	–0.54*	0.21	.01	457
	Lesson structure and facilitation	–0.71**	0.24	<.01	457
	Language and discourse	–0.57*	0.25	.02	457
Average baseline math score					
English language arts					
Framework for Teaching	Lesson structure and facilitation	0.28**	0.12	.02	457
Math					
Mathematical Quality of Instruction	Lesson structure and facilitation	–0.39*	0.16	.02	355
Percentage of students who are eligible for the federal school lunch program					
Math					
Mathematical Quality of Instruction	Feedback and assessment	1.02*	0.51	<.05	313

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Sample includes students who were randomly assigned to a classroom in year 2 of the Measures of Effective Teaching project. Reported results show the point estimates from regressing the standardized scores of a teacher observation instrument (overall or dimension specific), by subject, on the average classroom characteristics for students who were randomly assigned to a classroom. The point estimate is the coefficient on the student characteristic variable. Robust standard errors are reported. See table C7 in appendix C for complete results.

Source: Authors' analysis of Measures of Effective Teaching data.

classroom (23 percent of the 39 total regressions) and average baseline English language arts score (16 percent of the 39 total regressions) than for average baseline math score (5 percent of the 39 total regressions) or the percentage of students who are eligible for the federal school lunch program (<5 percent of the 39 total regressions). Third, significant effects were more often observed for the Framework for Teaching instrument (25 percent of the 48 total regressions) than for the Mathematical Quality of Instruction instrument (15 percent of the 20 total regressions), the Classroom Assessment Scoring System (<5 percent of the 64 total regressions), and Protocol for Language Arts Teaching Observations instruments (<5 percent of the 24 total regressions). Averaging instruments' results across dimension scores confirms that the strongest and most consistent relationships between teacher observation scores and

classroom composition were observed for the impact of the percentage of students who are racial/ethnic minority students, followed by average baseline English language arts scores, on observation scores from the Framework for Teaching instrument when scored on English language arts classroom instruction (see table C8 in appendix C).

The results indicate that assignment to a classroom with a larger percentage of racial/ethnic minority students or to one with a larger percentage of lower-achieving students reduces teachers' Framework for Teaching and Classroom Assessment Scoring System scores, but only in English language arts classrooms. Specifically, the findings for English language arts classes reveal significant relationships between classroom composition and observation scores for two instruments: Framework for Teaching and Classroom Assessment Scoring System. For English language arts classrooms, the percentage of students who are racial/ethnic minority students in the classroom and the classroom-average prior-year student test scores in English language arts both show statistically significant relationships with overall scores on the Framework for Teaching instrument and with dimension scores in four or five dimensions of instructional practice, respectively. Moreover, the classroom-average prior-year student test score in math was statistically significantly related to a Framework for Teaching dimension score in one dimension of instructional practice (lesson structure and facilitation). Similar patterns emerged for the Classroom Assessment Scoring System instrument for English language arts classrooms, but only with the percentage of students who are racial/ethnic minority students in the classroom and not classroom-average prior-year test scores. Specifically, the percentage of students who are racial/ethnic minority students in the classroom shows a statistically significant relationship with Classroom Assessment Scoring System dimension scores in two dimensions of instructional practice and with the overall Classroom Assessment Scoring System instrument score.

Assignment to a classroom with a larger percentage of racial/ethnic minority students or to one with a larger percentage of lower-achieving students reduces teachers' Framework for Teaching and Classroom Assessment Scoring System scores, but only in English language arts classrooms. The English language arts findings across instruments do not suggest that classroom composition affects any particular dimensions of instructional practice more consistently than others

The magnitudes of the effects of classroom composition are important to consider alongside statistical significance. By using the point estimates from table 6, one can calculate how a change in the composition of students in a classroom is expected to affect the median teacher's observation score. For example, increasing the percentage of students who are racial/ethnic minority students by 25 percentage points in an assigned English language arts classroom is expected to reduce the median teacher's score on the Classroom Assessment Scoring System instrument from the 50th percentile to approximately the 43rd percentile. Assignment to an English language arts class where the average prior-year English language arts test score is a quarter of a standard deviation below average would drop the median teacher's Framework for Teaching observation score from the 50th percentile to roughly the 45th percentile. As previously noted, the deviation in student composition between actual and randomly assigned classroom groups imply that the effects of student characteristics are somewhat larger than those reported here.

The results provide no evidence that classroom composition affects Protocol for Language Arts Teaching Observations scores (which covers only English language arts classes): none of the 24 Protocol for Language Arts Teaching Observations regressions produced a statistically significant finding.[5] In addition, the English language arts findings across instruments do not suggest that classroom composition affects any particular dimensions of instructional practice more consistently than others.

Unlike the results for English language arts classrooms, the results for math classrooms indicate that student composition does not consistently affect observation scores in math

classes. Only 3 of the 18 significant results are for math classes—no more than would be expected by chance; and none of those three results is for a composite observation score.

There are at least two possible explanations for the finding that student characteristics affect observation scores in English language arts classrooms. First, if the observation instruments are not sensitive to different kinds of instruction needed by different kinds of students, they might unfairly give lower scores to teachers of racial/ethnic minority or lower-achieving students who appropriately alter their instruction to meet student needs. In this case differences in scores would indicate a systematic bias in the observation instrument for the purpose of evaluating teacher instruction. But a second, alternative explanation is that teachers are providing less-effective instruction to non-White or low-achieving students. This might occur, for example, if teachers are subject to implicit biases that cause them to lower their expectations for such students. If changes in instructional practice are driven by systematic biases among teachers, differences in scores would indicate a systematic teaching problem rather than a bias in the observation instrument. The analysis cannot differentiate between these two potential explanations for the observed effect.

If the observation instruments are not sensitive to different kinds of instruction needed by different kinds of students, they might unfairly give lower scores to teachers of racial/ethnic minority or lower-achieving students who appropriately alter their instruction to meet student needs; however, it might be that teachers are inappropriately altering their instruction in the mistaken belief that different kinds of students need different kinds of instruction

Implications of the study findings

Seven of 10 dimensions of instructional practice were common across all five instruments, demonstrating the conceptual consistency of large parts of the different instruments. This finding is consistent with previous research that found strong correlations in scores for classroom observations rated on multiple instruments (Kane & Staiger, 2012).

At the same time the instruments differ in how many and which specific elements they measure within a given dimension of instruction. Some subdimensions are particularly rare among the instruments examined, such as family engagement, student perseverance, or teacher integrity with colleagues and parents (all captured only by the Framework for Teaching instrument). On average the Framework for Teaching instrument tends to provide the most coverage of elements within a given dimension, suggesting that it may offer a more comprehensive assessment of instructional practice than other instruments. When selecting among instruments, districts and states that prioritize particular dimensions of instructional practice should consider which instruments provide the best coverage of those dimensions, or which measure specific components of interest.

Across observation instruments classroom management is the dimension that is most strongly and consistently predictive of teachers' value-added scores. Districts and states that want observation scores, as measured by widely used instruments, to be more strongly related to teachers' contributions to student achievement growth might choose to place more weight on classroom management measures than other measures.

How the race/ethnicity, prior achievement level, or socioeconomic composition of students in the classroom affects observation scores varies by subject and by instrument. English language arts classes are more susceptible to classroom composition effects than are math classes: for two of the three instruments used to score English language arts instruction (Framework for Teaching and Classroom Assessment Scoring System), teachers with a larger percentage of racial/ethnic minority students tended to receive lower observation scores; a similar effect was observed for the Framework for Teaching instrument for teachers with lower-achieving students. The study finds no evidence that Protocol for Language

Arts Teaching Observations scores are related to the composition of students in English language arts classrooms, and little indication that observation scores are sensitive to student characteristics in math classes (particularly for Framework for Teaching and Classroom Assessment Scoring System scores, though less certainly for Mathematical Quality of Instruction scores).

It is possible that observation scores are sensitive to classroom composition in English language arts classes but not math classes due to differences in instructional styles across subjects. For example, if math instruction tends to take a more prescribed approach than English language arts instruction—for instance, by using common formats to structure lessons or pose questions to students—it might be easier for observers to reliably evaluate and to avoid rater bias or error that is correlated with student characteristics. Districts and states may wish to consider choosing one of the instruments for which scores do not appear to depend on classroom composition. This recommendation cannot be definitive because it is also possible that the classroom composition effects represent real differences in teaching practice rather than a bias in the instruments.

If math instruction tends to take a more prescribed approach than English language arts instruction—for instance, by using common formats to structure lessons or pose questions to students—it might be easier for observers to reliably evaluate and to avoid rater bias or error that is correlated with student characteristics

Nevertheless, the results can inform district and state efforts to reduce the likelihood of bias by helping them target limited resources toward the classrooms that are potentially most susceptible. For example, when training raters to conduct classroom observations, agencies might spend additional time honing raters' ability to reliably apply standards in the context of English language arts instruction.

Limitations of the study

One limitation of the study is that the observations were not conducted in circumstances resembling real teacher evaluations. No stakes were attached to the observation scores, which did not inform teachers' formal evaluations. Moreover, observations were conducted by trained and certified observers via videotape rather than by teachers' own principals during live classroom instruction. Relationships between teachers' observation scores and value-added scores might differ in the context of formal evaluations by school administrators with consequences for teachers. Moreover, the Measures of Effective Teaching project examined a convenience sample of six districts, and school and teacher volunteers within those districts, so the data are not representative of any district or education system. Relationships between observation scores and value-added scores, and the impact of classroom composition on observation scores, may differ for other educational contexts or populations of teachers.

In addition, the study statistically adjusted for measurement error in teachers' value-added scores but not their observation instrument scores. That is, the estimated correlations provide information about the relationship between teachers' value-added scores and their measured observation scores, but not actual performance, in different dimensions of instructional practice. The findings are thus useful for drawing conclusions about properties of various existing observation instrument measures, but they are more limited for drawing conclusions about true instructional practice. The relatively stronger and more consistent relationship to teachers' value-added scores for classroom management versus observation scores in other dimensions of instructional practice does not necessarily mean that classroom management is more important for student learning than other aspects of instruction. If it is easier to precisely measure performance in some dimensions of instructional

practice—such as classroom management—than others—such as student engagement, estimated correlations will not provide a fair comparison of the predictive power of different instructional practices. Caution should be used when interpreting findings in this way.

Another limitation is ambiguity about the underlying reason why the characteristics of students in the classroom affect teachers' observation scores. The study found that being assigned a larger percentage of racial/ethnic minority or lower-achieving students reduces English language arts teachers' observation scores on Classroom Assessment Scoring System and Framework for Teaching instruments, and the randomization of teachers to classes ensures that this relationship is causal. However, the study cannot distinguish what process leads to this effect which, as previously noted, may be due to two different processes. The first possibility is that the instruments are implicitly biased against teachers who have a larger percentage of racial/ethnic minority or lower-achieving students, by partially rating teachers on characteristics of the students they teach, rather than rating teachers only on their instructional practices. This would suggest a problem with using these instruments to evaluate individual teacher performance in English language arts classes. The second possibility is that English language arts teachers use less-effective teaching practices with racial/ethnic minority or lower-achieving students. This would suggest a teaching problem, rather than a problem with the instruments.

While the analysis identifies a causal effect of student characteristics on some observation scores, it cannot isolate the underlying mechanism that explains the effect

Thus, while the analysis identifies a causal effect of student characteristics on some observation scores, it cannot isolate the underlying mechanism that explains the effect. Care should be used when interpreting this finding's implications. For instance, adjusting teacher observation scores by student characteristics would improve their validity as measures of teacher performance only if the relationship between classroom composition and teacher observation scores is due to bias in the observation instrument. If observation scores instead vary by classroom composition because teachers are less effective with certain types of students, then controlling for student characteristics would be counterproductive for improving student learning and ensuring equitable access to effective instruction. Nonetheless, the fact that Protocol for Language Arts Teaching Observations scores are not affected by the characteristics of students in the classroom suggests reason for concern that an implicit bias may be inherent in the Classroom Assessment Scoring System and Framework for Teaching instruments when used in English language arts classes.

Appendix A. Detailed study methodology

The study used two sources of data: five widely used teacher observation instruments and data from the Measures of Effective Teaching (MET) longitudinal database on the 2009/10 and 2010/11 school years.[6] The following subsections provide a summary of the methodology used to answer each research question.

Instrument content analysis of elements of instructional practice

The goal of the content analysis was to identify common categories of instructional practice across five observation instruments: the Classroom Assessment Scoring System (CLASS), the Framework for Teaching (FFT), the Protocol for Language Arts Teaching Observations (PLATO), the Mathematical Quality of Instruction (MQI), and the UTeach Observational Protocol (UTOP). The rubric for each instrument is pre-organized into categories of instructional practice, but these instruments cannot be compared directly because of two key differences in how they are constructed (table A1):

- ***The instruments differ in the detail used to define predetermined categories of instructional practice.*** For example, both the CLASS and FFT instruments identify monitoring of student behavior as a category of instructional practice, and both instruments nest this practice under a broader category of behavior management, which is further nested under classroom organization/environment (see table A1, example 1). Yet the CLASS instrument includes an additional subdimension (proactive) between the monitoring category and behavior management category, while the FFT instrument does not. Moreover, the FFT instrument identifies an additional subelement below its monitoring category (teacher awareness of student conduct), while the CLASS instrument does not.
- ***The instruments differ in how they organize overlapping subdimensions into higher-level groups.*** For example, both the CLASS and FFT instruments identify respectful language and listening as subelement indicators (see table A1, example 2). Yet the CLASS instrument categorizes these concepts under a domain titled emotional support, while the FFT categorizes them under a domain titled classroom environment.

In short, while each instrument rubric is pre-organized into categories of instructional practice, the instruments vary in how fine grained these categories are and, in some cases, in how subcategories are grouped into larger categories. Because of these differences, the study team derived dimensions and subdimensions from textual analysis rather than choosing from among the predetermined set defined by instrument developers. That is, the study team used an inductive approach to identify common categories of instructional practice across instruments.

The instrument content analysis proceeded in three stages: inductive coding, verification literature review, and focused coding. The purpose of the first stage of analysis was to derive empirically determined dimensions and subdimensions of instructional practice. The second stage of analysis sought to verify whether the empirically determined dimensions and subdimensions were consistent with the categories, definitions, and underlying theory of instruction that informed each instrument's development. In the third stage of analysis, the goal was to assess the reliability of the dimensions and subdimensions identified in the

Table A1. Illustration of differences across instruments in developer-defined categories of instructional practice

Level of category	Classroom Assessment Scoring System (CLASS)	Framework for Teaching (FFT)
Example 1: Monitoring of student behavior		
Level 1: Domain	Classroom organization	Classroom environment
Level 2: Subdomain	Behavior management	Managing student behavior
Level 3: Subdomain element	Proactive	Monitoring of student behavior
Level 4: Subelement indicator	Monitoring	Teacher awareness of student conduct
Example 2: Respectful language and active listening		
Level 1: Domain	Emotional support	Classroom environment
Level 2: Subdomain	Positive climate	Creating an environment of respect and rapport
Level 3: Subdomain element	Respect	Teacher interactions with students, including both words and actions
Level 4: Subelement indicators	Respectful language Listening to each other	Respectful talk, active listening, and turn-taking

Source: Pianta, Hamre, & Mintz, 2012; Danielson, 2014.

first stage of analysis, to refine the codes and definitions, and to assess the degree to which each dimension and subdimension was represented across the five instruments.

Stage 1: Inductive coding. For the inductive coding one member of the study team applied descriptive codes to each instrument's scoring rubric and supporting documentation (table A2) to identify descriptive categories. The five observation protocols and supporting documentation were inductively coded by summarizing text and applying descriptive codes—which group text into descriptive categories that closely reflect the original text—followed by interpretive codes—which group descriptive codes into more inferential, higher-level categories (Miles & Huberman, 1994). From these codes the member of the study team developed a focused coding scheme to be used in a second round of coding. The coding scheme included 10 interpretive codes intended to represent higher-level dimensions of instructional practice and 37 interpretive codes categorized as subdimensions of instructional practice (see table C2 in appendix C).

The member of the study team then examined relationships among the descriptive categories and organized them into a hierarchical structure with more specific elements using interpretive codes, such as clear learning objectives, nested within higher-level categories such as lesson structure and facilitation.

Stage 2: Verification literature review. As a verification check, a second member of the study team conducted a review of relevant literature for each instrument to confirm the theoretical and empirical basis for the identified dimensions and subdimensions of instructional practice. The purpose of the verification check was to assess whether the dimensions and subdimensions derived from the analysis were consistent with developers' categories, definitions, and underlying theories of instruction that informed each instrument's development. Specifically, the member of the study team who did not participate in the inductive (or deductive) coding independently reviewed content on each instrument's design,

Table A2. Content analysis data sources

Instrument	Data sources
Classroom Assessment Scoring System (CLASS)	Pianta et al., 2012
Framework for Teaching (FFT)	Danielson, 2015 Danielson, 2014
Protocol for Language Arts Teaching Observations (PLATO)	Grossman & Greenberg, n.d.[a]
Mathematical Quality of Instruction (MQI)	Hill, 2014
UTeach Observational Protocol (UTOP)	UTeach, 2014

a. The study copy of this protocol was not marked with a date, and the study team was unable to obtain a revised version from the developers or the Stanford Graduate School of Education.

purpose, and theoretical underpinnings. Such sources included relevant published articles by the instrument's creator, an instrument's training guide or manual, and information available on an instrument's official website. After reviewing these sources, the member of the study team then re-examined the coding structure, particularly how codes were grouped in relation to one another, and made any appropriate adjustments to the coding structure. The resulting structure served as a coding scheme for focused coding in the next stage of analysis (see table C2 in appendix C).

Stage 3: Focused coding. For the focused coding two members of the study team deductively applied codes to the observation instrument scoring rubrics, using ATLAS.ti qualitative software and the focused coding scheme developed in the previous analysis stage. After attending a one-hour training session, the members of the study team independently coded the rubrics in three sets: CLASS and FFT rubrics, MQI rubric, and PLATO and UTOP rubrics. After coding each set of rubrics, the members of the study team met to compare codes, discuss any disagreements, reach consensus on which descriptive and interpretive codes apply to each phrase, and adjust code definitions or the coding scheme as necessary. The members of the study team coded the five rubrics with an overall initial inter-rater reliability of 65.4 percent (applying the same code to 159 of 243 total phrases coded), achieving 100 percent agreement after resolving the 84 discordantly coded phrases in code review meetings.

Correlation analysis of teacher observation scores and value-added scores

The study team examined correlations between teachers' observation instrument dimension scores and teachers' value-added scores. The study team first calculated teachers' contributions to student achievement gains in English language arts and math, as measured by state-administered tests that varied by participating district, each of which was located in a different state. These data were used to estimate teacher value-added models.

Teacher value-added model estimation. To examine correlations between teachers' value-added scores and scores on teacher observation instruments, the study team calculated value added using the following equation:[7]

$$Y_{i,t,c} = Y_{i,t-1}\lambda + X_{i,t}\beta + \bar{X}_{i,t,c}\gamma + D_{i,t}\delta + M_t\tau + e_{i,t,c} \quad \text{(A1)}$$

In equation A1 the variable $Y_{i,t,c}$ represents the standardized English language arts or math test score of student i in year t and classroom c.[8] Each student's prior-year scores on state

assessments in English language arts and math are included as control variables indicated by the vector $Y_{i,t-1}$. $X_{i,t}$ is a vector of six student background characteristics (gender, race/ethnicity, eligibility for the federal school lunch program, special education status, English language learner status, and gifted status). $\bar{X}_{i,t,c}$ is a vector of three classroomwide average student characteristics (classroomwide mean English language arts and math baseline test score, number of students in the classroom, and mean age of students), to account for peer effects on student achievement.[9] All variables in the model were centered at the mean of the analytic sample. $D_{i,t}$ is a set of teacher fixed effects, M_t is a vector of school year indicator variables (for models including data on more than one school year) and grade-level indicator variables (for those including data on more than one grade level), and $e_{i,t,c}$ is an error term. Because the model excludes a constant term, δ is a vector of teacher value-added coefficients, expressing how each teacher's contribution to student learning compares with what is expected for the average teacher (see Rotz, Johnson, & Gill, 2014). In total the 30 models described in table A3 were estimated.

Equation A1 was estimated separately by grade-level groups (4–5, 6–8, or 9),[10] subject (English language arts or math), and district, as each district was located in a different state and thus used a different assessment. Finally, teachers' value-added scores were estimated on a different group of students than those students contributing to their teacher observation scores. The grade 6–9 models included up to two years of data per teacher, estimating average teacher contributions to student achievement across the 2009/10 and 2010/11 school years. Because most teachers in grades 6–9 taught multiple classes per year, these samples included only students in those classes not scored by one of the observation instruments of interest (CLASS, FFT, PLATO, MQI, or UTOP). For grades 4–5, most teachers taught only one class per year. Therefore, models were estimated separately by school year for all students assigned to a grade 4 or grade 5 teacher in that year.

To maximize sample size and use all students with an observed outcome to estimate a teacher's value-added score, the study team imputed missing baseline control variables, represented in equation A1 by the vectors, $Y_{i,t-1}$, $X_{i,t}$, and $\bar{X}_{i,t,c}$ (see appendix B for a description of imputation procedures and sample characteristics, with and without imputed values).

Table A3. Description of teacher value-added models, by sample

Outcome	Grades 4–5 (year 1)	Grades 4–5 (year 2)	Grades 6–8 (all years)	Grade 9 (all years)
State English language arts assessment score	District 1 District 2 District 3 District 5	District 1 District 2 District 3 District 5	District 1 District 3 District 4 District 5 District 6	District 1 District 2
State math assessment score	District 1 District 2 District 3 District 5	District 1 District 2 District 3 District 5	District 1 District 3 District 4 District 5 District 6	District 1 District 2
Number of models	16		10	4

Note: State assessments varied by state and district in the Measures of Effective Teaching project. Project participants did not include any grade 4–5 teachers from district 6, any grade 6–8 teachers from district 2, or any grade 9 teachers from districts 3–6. Grade 4–5 teachers from district 4 were excluded from these analyses because data on eligibility for the federal school lunch program were missing for all students.

Source: Authors' analysis of Measures of Effective Teaching data.

Construction of teachers' observation instrument dimension scores. The study team used teachers' observation instrument scores from the MET project to derive dimension scores. The study team constructed dimension scores from pre-existing subscores, which either combine multiple scores into a single score or represent a single score for a group of practices, as defined by the instrument developer.[11] The MET database includes subscores only for portions of each rubric that can be scored from classroom videos alone. As a result subscores for some components identified in the content analysis, such as the teacher professionalism component of the FFT rubric, were not available in the MET database and were thus excluded from this analysis. To construct dimension scores, the study team first grouped the available pre-existing subscores, as shown in table A4, into the 10 dimensions of instructional practice identified in the content analysis.

Table A4. Dimension scores comprising subscores, by observation instrument

Dimension and instrument	Instrument subscores included in dimension score
Supportive learning environment	
CLASS	Positive climate, negative climate, teacher sensitivity
FFT	Creating environment of respect and rapport, establishing a culture of learning
UTOP	Collegiality among students; attention to access, equity, and diversity
Student focus	
CLASS	Regard for student perspectives
Classroom management	
CLASS	Behavior management, productivity
FFT	Managing student behavior, managing classroom procedures
PLATO	Behavior management, time management
UTOP	Classroom management, majority of students on task
Active student participation in class activities	
CLASS	Student engagement[a]
Student intellectual engagement with content	
CLASS	Analysis and problem solving
PLATO	Intellectual challenge
MQI	Student participation in meaning-making and reasoning
UTOP	Investigation/problem-based approach, intellectual engagement with key ideas
Lesson structure and facilitation	
CLASS	Instructional learning formats, student engagement[a]
FFT	Communicating with students, engaging students in learning
PLATO	Explicit strategy use and instruction
MQI	Classroom work connected to mathematics
UTOP	Lesson organization, student generation of ideas/questions, appropriate resources, involvement of all students, allocation of time, structures for student engagement
Content understanding	
CLASS	Content understanding
PLATO	Representations of content, modeling
MQI	Richness of mathematics, errors and imprecision, explicitness and thoroughness in content presentation
UTOP	Significance of content, explicitness of content importance, teacher knowledge and content fluency, accuracy of teacher written content, use of content abstraction and representation, connections to other disciplines, relevance to history and current events

(continued)

Table A4. Dimension scores comprising subscores, by observation instrument
(continued)

Dimension and instrument	Instrument subscores included in dimension score
Language and discourse	
CLASS	Instructional dialogue
FFT	Using questioning and discussion techniques
PLATO	Classroom discourse
UTOP	Questioning strategies
Feedback and assessment	
CLASS	Quality of feedback
FFT	Using assessment in instruction
MQI	Working with students and mathematics
UTOP	Use of formative assessments
Teacher professionalism	
UTOP	Lesson reflection

CLASS is Classroom Assessment Scoring System. FFT is Framework for Teaching. PLATO is Protocol for Language Arts Teaching Observations. MQI is Mathematical Quality of Instruction. UTOP is UTeach Observational Protocol.

a. The analysis focused on the dimensions for which the Measures of Effective Teaching data included subscores from at least three instruments; however, to use all possible information, the study team used the CLASS student engagement subscore as a proxy measure for teacher efforts to engage students, under the lesson structure and facilitation dimension.

Source: Instrument subscores from Measures of Effective Teaching data and author's content analysis of observation instrument rubrics.

To enable comparison across instruments, the analysis focused on the seven dimensions for which the MET data included subscores from more than one instrument: supportive learning environment, classroom management, student intellectual engagement with content, lesson structure and facilitation, content understanding, language and discourse, and feedback and assessment.[12]

The study team calculated instrument-specific dimension scores as the mean of each instrument's relevant subscores within a given dimension. For example, the positive climate, negative climate, and teacher sensitivity subscores were averaged to yield a supportive learning environment dimension score for the CLASS instrument. For a given dimension of instructional practice, three to five instrument-specific dimension scores were available in the MET data (table A5).[13]

The study team calculated cross-instrument dimension scores from all instrument-specific scores for a given dimension, calculated separately by subject (English language arts or math). For example, to create an overall dimension score for supportive learning environment, the study team first standardized the instrument-specific dimension scores for the CLASS, FFT, and UTOP instruments, so that each had a mean of 0 and standard deviation of 1. The overall supportive learning environment score for English language arts classes was then calculated as the mean of the standardized versions of the CLASS and FFT dimension scores. The overall score for math classes was calculated as the mean of the standardized CLASS, FFT, and UTOP dimension scores.

Table A5. Dimension scores available in Measures of Effective Teaching data, for more than one teacher observation instrument

Dimension	Classroom Assessment Scoring System (CLASS)	Framework for Teaching (FFT)	Protocol for Language Arts Teaching Observations (PLATO)	Mathematical Quality of Instruction (MQI)	UTeach Observational Protocol (UTOP)
Supportive learning environment	✔	✔			✔
Classroom management	✔	✔	✔		✔
Student intellectual engagement with content	✔		✔	✔	✔
Lesson structure and facilitation	✔	✔	✔	✔	✔
Content understanding	✔		✔	✔	✔
Language and discourse	✔	✔	✔		✔
Feedback and assessment	✔	✔		✔	✔

Source: Author's quantitative analysis of instrument scores using Measures of Effective Teaching data.

Correlation with underlying value added. The study team then correlated teachers' value-added scores with teachers' observation instrument scores, expressed as instrument-specific or cross-instrument dimension scores.

The study estimated teacher-level correlations separately by subject (English language arts or math) but pooled across districts and, for grades 6–9, pooled across grade levels. The teacher value-added scores were first standardized by district, subject, grade-level grouping (4–5, 6–8, and 9), and year (if applicable) to have a mean of zero and standard deviation of one within subgroups. The instrument scores were also standardized by dimension score, instrument type, and subject (English language arts or math class).

The study team estimated correlations between teachers' value-added scores (each of four subject-by-year scores for grade 4–5 teachers and each of two subject-specific scores for grade 6–9 teachers) and teachers' observation dimension scores (instrument-specific or cross-instrument) in each of seven dimensions of instructional practice (see table A6).

The purpose of estimating these correlations was to assess the degree to which teachers' observation scores in various dimensions of instructional practice are related to teachers' value-added scores. However, the study uses observation scores and value-added scores that are measured with error. A random element associated with a particular type of student (such as an unusually disruptive student) could affect both value-added scores and observation scores for that classroom, inflating the correlation between the two measures for reasons unrelated to the teacher's true effectiveness (see Chaplin et al., 2014). Thus, to ensure that correlations between teachers' value-added scores and teacher observation dimension scores were not inflated by random events for both scores in the same direction, the study correlated observation scores observed with one group of students with value-added scores estimated on a different group of students—all those taught by the teacher who were not present for a classroom lesson observed by MET. For grade 6–9 teachers, who were largely subject-matter specialists teaching multiple classes of students per school year, teachers' value added was estimated only for students in classrooms not

Table A6. Teacher value-added scores and observation instrument dimension scores, by subject

Grade level and dimension	English language arts	Math
Teacher value-added scores		
Grades 4–5	State English language arts, year 1 State English language arts, year 2	State math, year 1 State math, year 2
Grades 6–9	State English language arts, both years	State math, both years
Teacher observation instrument dimension scores		
Supportive learning environment	Overall score (across instruments) CLASS FFT	Overall score (across instruments) CLASS FFT UTOP
Classroom management	Overall score (across instruments) CLASS FFT PLATO	Overall score (across instruments) CLASS FFT UTOP
Lesson structure and facilitation	Overall score (across instruments) CLASS FFT PLATO	Overall score (across instruments) CLASS FFT MQI UTOP
Intellectual engagement	Overall score (across instruments) CLASS PLATO	Overall score (across instruments) CLASS MQI UTOP
Feedback and assessment	Overall score (across instruments) CLASS FFT	Overall score (across instruments) CLASS FFT MQI UTOP
Language and discourse	Overall score (across instruments) CLASS FFT PLATO	Overall score (across instruments) CLASS FFT UTOP
Content understanding	Overall score (across instruments) CLASS PLATO	Overall score (across instruments) CLASS MQI UTOP

CLASS is Classroom Assessment Scoring System. FFT is Framework for Teaching. PLATO is Protocol for Language Arts Teaching Observations. MQI is Mathematical Quality of Instruction. UTOP is UTeach Observational Protocol.

Note: State assessments varied by state and district in the study.

Source: Authors' analysis of Measures of Effective Teaching data.

scored by any of the observation instruments of focus (that is, CLASS, FFT, PLATO, MQI, or UTOP). More specifically, the correlation was estimated as:

$$\rho(VA_{i,s}, OS_{i,s'}) = \frac{Cov(VA_{i,s}, OS_{i,s'})}{\sqrt{(Var\ VA_{i,s}) * Var(OS_{i,s'})}} \quad \text{(A2)}$$

In equation A2, $VA_{i,s}$ represents the average value-added score for teacher i to learning among group of students, s. The variable $OS_{i,s'}$ represents the observation dimension score of the same teacher, i, observed while teaching group of students s', a group of students other than s.

Finally, to correct for year-to-year changes in teachers' value-added scores, the study adjusted the correlations estimated in equation A2 for the "between-year correlation in achievement gains" (Kane & Staiger, 2012, p. 45). This approach was used because the goal was to estimate the covariation of teacher observation scores with the component of their teaching quality that does not change over time.[14] Using this approach, the study team estimated the correlation between observation dimension scores and a more generalized teacher value-added measure as:

$$\rho(VA_i, OS_{i,s'}) = \frac{\rho(VA_{i,s}, OS_{i,s'})}{\sqrt{\rho(VA_{i2009-10}, VA_{i2010-11})}} \quad \text{(A3)}$$

In equation A3, VA_i represents teacher *i*'s more consistent (that is, time-invariant) value added, not specific to the idiosyncrasies of a particular school year or group of students. The numerator is the correlation coefficient estimated using equation A2. The denominator is the square root of the Pearson's correlation between value-added scores estimated for two school years, 2009/10 and 2010/11. For grades 4–5 the between-year correlation in estimates of teacher value-added was 0.38 for English language arts and 0.61 for math. For grades 6–9 the correlation was 0.32 for English language arts and 0.20 for math.

Correlation analysis of teacher observation scores and student characteristics

Sample construction. To test for effects of classroom composition on each teacher observation instrument, the study team took advantage of the subsample of teachers who were randomized to a classroom of students in year 2 of the MET project. Randomization was done by block, with each block representing a group of teachers within a school who were teaching the same subject to the same grade range of students. The study team conducted an intent-to-treat design where classroom characteristics were calculated based on the students that were supposed to be in each teacher's randomly assigned classroom irrespective of whether they actually ended up being in that classroom.

Furthermore, the MET dataset only follows teachers who volunteered to be in the study. There is no data on students who were randomized to be in a classroom with a MET project teacher, but actually enrolled in a classroom with a teacher who was not in the MET project. This noncompliance resulted in limited data on the number of students who were randomly assigned to classrooms. To reduce the chance of outlier students biasing classroom composition characteristics, all classes within a randomization block were dropped if any classroom in the block included data on fewer than 10 students. In addition, to reduce the chance of bias due to randomly assigned students not being in a MET classroom, blocks were dropped if there was greater than a 50 percent difference in the number of assigned students with data between the largest and smallest classrooms in the block. For example, if a randomization block had six classrooms and data on 12 students in the smallest classroom but 20 students in the largest classroom (that is, data on 166 percent more students than in the smallest classroom), the study would exclude all six classes from the analysis due to the discrepancy in number of students per classroom. If another randomization block had four classrooms and data on only 8 students for one of those classrooms, the study would exclude all four classrooms from the analysis.

Table A7. Sample definition for analysis of observation scores and student characteristics

Sample	Condition	Number of class sections
A	All randomized English language arts or math class sections in the study	1,252
B	All teachers with a randomized class section and with data on at least 1 student randomly assigned	1,181
C	Only teachers where all teachers in a randomization block have at least 10 students randomly assigned	726
D	Only teachers where all teachers in a randomization block have no greater than a 50 percent disparity in the number of students randomly assigned	754
E	When both conditions C and D are met	671

Source: Authors' analysis of Measures of Effective Teaching data.

Finally, this analysis focuses only on English language arts and math classes, so all grade 9 biology classes were excluded from the analysis. The study's analytic sample for this analysis is defined in sample E in table A7.

Of the 671 teachers in the analytic sample, there was outcome data (teacher observation score) for 662 of those teachers. Nine teachers (about 1 percent of the analytic sample) did not end up teaching a class within the MET project sample.

Random assignment verification check. The study team conducted a verification check on randomization for the study's analytic sample to confirm that the absence of data on students who left the study sample did not undermine the randomization. The study team regressed year 1 standardized teacher observation scores on year 2 randomly assigned classroom characteristics.[15] There should be no relationship between a teacher's observation scores in year 1 of the MET project with the classroom characteristics for a different group of students to which the teacher was randomly assigned in year 2 of the study. The regression model is:

$$Y_{i,1} = \alpha + \beta X_{i,2} + \gamma Z_{i,2} + e_i \qquad \text{(A4)}$$

In equation A4, $Y_{i,1}$ is the average year 1 standardized observation score (overall or subdimension) for teacher i on a given instrument.[16] The variable $X_{i,2}$ is one of the four classroom composition measures for teacher i's randomly assigned classroom in year 2 (percentage of students who are racial/ethnic minority students, percentage of students who are eligible for the federal school lunch program, prior-year English language arts test scores, and prior-year math test scores).[17] The component $Z_{i,2}$ represents a vector of dummy indicators for randomization blocks that teacher i was assigned to in year 2 of the study. The error term is indicated by e_i. For the CLASS and FFT instruments, which were used to score both English language arts and math classes in the MET project, the study team estimated separate models by subject (English language arts versus math). Robust standard errors were estimated.

Of the 156 regressions run using equation A4, about 4 percent (seven models) yielded statistically significant relationships between randomly assigned classroom characteristics in year 2 and teacher observation scores in year 1, using a significance level of .05. This is

what is expected due to chance alone. This suggests that the missing data did not undermine the success of the random assignment.

Estimation of classroom composition impact. To estimate the effect of classroom characteristics on observation instrument scores (overall or dimension specific), the study team regressed year 2 standardized observation scores on year 2 randomly assigned classroom characteristics using the following equation:

$$Y_{i,2} = \alpha + \beta X_{i,2} + \gamma Z_{i,2} + e_i \quad \text{(A5)}$$

Equation A5 is identical to equation A4 except that the dependent variable is now the year 2 standardized observation scores, representing either an overall instrument score or a dimension score for an instrument. Robust standard errors are again estimated.

Appendix B. Imputation methodology for value-added model estimation

To maximize use of available data in the estimation of teachers' value-added scores, the study team imputed missing baseline control variables via a method used in other rigorous studies, single stochastic regression imputation (Tuttle et al., 2013). The imputation of missing baseline data was implemented with the ice command in Stata. Data were missing on 12 covariates measuring student background characteristics, 6 at the student level and 6 at the classroom level:

- Prior-year English language arts test score.
- Prior-year math test score.
- Gifted indicator.
- Special education indicator.
- Age.
- Eligibility for the federal school lunch program indicator.

No students were missing data on gender or race/ethnicity. For each variable, M, with missing observations, the study team estimated equation B1 using observed data on students *i* in years *t* and classrooms *c*.

$$M_{i,t,c} = \alpha + Y_{i,t-n},\lambda + X_{i,t}\beta + X_{i,t,c}\gamma + e_{i,t,c} \quad \text{(B1)}$$

Equation B1 includes the same student- and classroom-level covariates as those used in the value-added models (equation A1). Similarly, the vector $Y_{i,t-n},\lambda$ again represents students' prior state standardized test scores; however, in this case, the vector includes scores for one, two, and three years prior to year *t*, and for science, social studies, and writing, in addition to English language arts and math. For each covariate missing observations, we imputed missing values as the sum of the predicted value of $M_{i,t,c}$, calculated using the coefficients estimated when fitting equation B1, and a stochastic component randomly selected from the set of estimated model residuals ($e_{i,t,c}$). The study used an iterative procedure to allow for cases with multiple missing values. Specifically, the imputation equations were estimated in 10 cycles, first using data from cases with no missing observations then re-estimating each imputation model on an updated dataset including both complete-case data and (newly) imputed data (Lunt, 2011).

Tables B1 and B2 report summary statistics with and without imputed missing values, for the sample of students, classrooms, and teachers included in at least one analysis.

Table B1. Summary statistics for measures of student characteristics, with and without imputed missing values

Item	Mean	Standard deviation	Minimum	Maximum	Number of students
Sample with imputed data					
Prior-year English language arts test score	0.05	0.93	–3.97	3.59	117,703
Prior-year math score	0.05	0.93	–3.99	3.69	117,703
Gifted student	0.09	0.29	0	1	117,703
Male	0.50	0.50	0	1	117,703
Special education student	0.08	0.27	0	1	117,703
English language learner	0.13	0.34	0	1	117,703
Age (years)	12.09	1.88	6.97	20.40	117,703
Eligible for the federal school lunch program	0.56	0.50	0	1	117,703
White (non-Hispanic)	0.24	0.43	0	1	117,703
Hispanic	0.32	0.47	0	1	117,703
Black (non-Hispanic)	0.35	0.48	0	1	117,703
Asian (non-Hispanic)	0.06	0.25	0	1	117,703
American Indian (non-Hispanic)	0.00	0.06	0	1	117,703
Other race/ethnicity	0.02	0.15	0	1	117,703
Sample without imputed data (original data)					
Prior-year English language arts test score	0.08	0.93	–3.70	3.59	105,424
Prior-year math test score	0.08	0.94	–3.68	3.69	106,081
Gifted student	0.09	0.29	0	1	115,844
Male	0.50	0.50	0	1	117,703
Special education student	0.08	0.27	0	1	117,072
English language learner	0.13	0.34	0	1	117,703
Age (years)	12.09	1.88	6.97	20.40	117,416
Eligible for the federal school lunch program	0.57	0.49	0	1	100,869
White (non-Hispanic)	0.24	0.43	0	1	117,703
Hispanic	0.32	0.47	0	1	117,703
Black (non-Hispanic)	0.35	0.48	0	1	117,703
Asian (non-Hispanic)	0.06	0.25	0	1	117,703
American Indian (non-Hispanic)	0.00	0.06	0	1	117,703
Other race/ethnicity	0.02	0.15	0	1	117,703

Note: Sample includes the 117,703 students assigned to the 5,409 teachers included in any of the correlation analyses to assess the relationship between teachers' value-added and observation scores.

Source: Authors' analysis of Measures of Effective Teaching data.

Table B2. Summary statistics for measures of classroom characteristics, with and without imputed missing values

Item	Mean	Standard deviation	Minimum	Maximum	Number of students
Sample with imputed data					
Prior-year English language arts test score	0.02	0.60	–2.52	2.35	5,409
Prior-year math score	0.02	0.60	–2.73	1.93	5,409
Gifted student	0.09	0.17	–0.34	1	5,409
Male	0.50	0.13	0	1	5,409
Special education student	0.09	0.13	0	1	5,409
English language learner	0.14	0.18	0	1	5,409
Age (years)	12.01	1.86	8.09	17.58	5,409
Eligible for the federal school lunch program	0.56	0.30	–0.21	1.28	5,409
White (non-Hispanic)	0.25	0.27	0	1	5,409
Hispanic	0.32	0.28	0	1	5,409
Black (non-Hispanic)	0.35	0.32	0	1	5,409
Asian (non-Hispanic)	0.06	0.11	0	1	5,409
American Indian (non-Hispanic)	0.03	0.04	0	0.31	5,409
Other race/ethnicity	24.67	6.97	2	71	5,409
Sample without imputed data (original data)					
Prior-year English language arts test score	0.02	0.60	–2.25	2.35	5,363
Prior-year math test score	0.02	0.61	–2.73	1.93	5,367
Gifted student	0.09	0.17	0	1	5,301
Male	0.5	0.13	0	1	5,409
Special education student	0.09	0.13	0	1	5,408
English language learner	0.14	0.18	0	1	5,409
Age (years)	12.01	1.86	8.09	17.58	5,409
Eligible for the federal school lunch program	0.58	0.30	0	1	4,470
White (non-Hispanic)	0.25	0.27	0	1	5,409
Hispanic	0.32	0.28	0	1	5,409
Black (non-Hispanic)	0.35	0.32	0	1	5,409
Asian (non-Hispanic)	0.06	0.11	0	1	5,409
American Indian (non-Hispanic)	0.03	0.04	0	0.31	5,409
Other race/ethnicity	24.67	6.97	2	71	5,409

Note: Sample includes the 5,409 teachers included in any of the correlation analyses to assess the relationship between teachers' value-added and observation scores.

Source: Authors' analysis of Measures of Effective Teaching data.

Appendix C. Supplementary results

This appendix presents supplementary results for the three main analyses presented in this report: the content analysis of observation instruments (tables C1–C3), the correlation analysis between teachers' observation scores and their value-added scores (tables C4–C6), and the regression analysis of how classroom composition affects teachers' observation scores (tables C7–C8). More specifically, it reports the following:

- Subdimensions rated by each observation instrument (table C1).
- Subdimension definitions derived through inductive coding (table C2).
- Examples of coded units for each dimension of instructional practice (table C3).
- Selected teacher value-added model results (table C4).
- Summary results by subject and grade level for the strength of relationships between teachers' observation scores and their value-added scores (table C5).
- Summary results by instrument for the strength and consistency of relationships between teachers' observation scores and their value-added scores (table C6).
- Complete results for the relationship between classroom composition and teachers' observation scores (table C7).
- Summary results by instrument for the strength and consistency of relationships between classroom composition and teachers' observation scores (table C8).

Table C1. Instrument content analysis: Subdimensions of instructional practice rated by each observation instrument

Subdimension	Classroom Assessment Scoring System (CLASS)	Framework for Teaching (FFT)	Protocol for Language Arts Teaching Observations (PLATO)	Mathematical Quality of Instruction (MQI)	UTeach Observational Protocol (UTOP)
Supportive learning environment					
Teacher awareness and responsiveness	✔	✔		✔	✔
Teacher–student positive energy and rapport	✔	✔	✔		✔
Inclusive class environment			✔		✔
Student ease in educational environment	✔	✔	✔		✔
High expectations for students		✔	✔		
Student focus					
Active listening and encouragement of student ideas	✔		✔	✔	✔
Student autonomy/leadership or productive group work	✔	✔	✔		✔
Teacher flexibility and tailoring to student needs	✔	✔	✔	✔	
Family engagement		✔			
Classroom management					
Behavior management	✔	✔	✔	✔	✔
Time management	✔	✔	✔	✔	✔
Active student participation in class activities					
Active student participation in class activities	✔		✔	✔	

(continued)

Table C1. Instrument content analysis: Subdimensions of instructional practice rated by each observation instrument *(continued)*

Subdimension	Classroom Assessment Scoring System (CLASS)	Framework for Teaching (FFT)	Protocol for Language Arts Teaching Observations (PLATO)	Mathematical Quality of Instruction (MQI)	UTeach Observational Protocol (UTOP)
Student intellectual engagement with content					
Cognitive challenge		✔	✔	✔	
Student connection questions or alternative ideas		✔	✔	✔	
Student explanation, prediction, or investigation	✔	✔	✔	✔	✔
Student perseverance		✔			
Lesson structure and facilitation					
Clear learning objectives	✔	✔	✔	✔	✔
Clear presentation, sequence, or effective pacing	✔	✔		✔	✔
Variety of strategies, materials, or efforts to engage students	✔	✔	✔	✔	✔
Effective classroom setup		✔			✔
Content understanding					
Explicit, in-depth explanation of concepts or procedures	✔	✔	✔	✔	✔
Rich and meaningful content	✔	✔	✔	✔	✔
Connections to real world or prior knowledge and experiences	✔		✔	✔	✔
Use of multiple representations or examples	✔		✔	✔	✔
Opportunities to practice applying concepts or procedures	✔		✔	✔	
Teacher content understanding and accuracy		✔	✔	✔	✔
Language and discourse					
In-depth, content-driven student discussion	✔		✔	✔	✔
Use of academic language		✔	✔	✔	
Discussion facilitation and cumulative exchanges	✔		✔		
Feedback and assessment					
Back and forth teacher–student feedback exchanges	✔		✔		
Scaffolding feedback	✔		✔		✔
Specific feedback	✔	✔	✔	✔	
Correcting student misconceptions	✔			✔	✔
Checking for student understanding		✔	✔	✔	✔
Assessment linked to objectives		✔			✔

(continued)

Table C1. Instrument content analysis: Subdimensions of instructional practice rated by each observation instrument *(continued)*

Subdimension	Classroom Assessment Scoring System (CLASS)	Framework for Teaching (FFT)	Protocol for Language Arts Teaching Observations (PLATO)	Mathematical Quality of Instruction (MQI)	UTeach Observational Protocol (UTOP)
Teacher professionalism					
Reflective teacher practice		✔			✔
Collaboration, leadership, and professional learning		✔			
Integrity with colleagues and parents		✔			

Note: See table C2 for the authors' subdimension definitions and table C3 for example instrument text coded for each dimension.

Source: Authors' content analysis of observation instrument rubrics.

Table C2. Instrument content analysis: Focused-coding scheme with definitions developed through inductive coding

Code	Definition
Supportive learning environment	
Teacher awareness and responsiveness	Degree to which the teacher demonstrates awareness of student needs or problems (such as by monitoring class for students who have questions or are confused), anticipates issues that students might have, offers assistance or support to individual students, is effective in addressing student problems, responds quickly to student needs, adjusts lesson speed or wait time, and recognizes students' out-of-school issues.
Teacher–student positive energy and rapport	Degree to which the teacher and students are physically close to each other, engage in social conversation, display warm/supportive interactions, offer each other praise and encouragement, listen to each other, cooperate, use names and respectful language/tone, and smile or laugh. Degree to which the classroom is free of anger, harsh voices, physical aggression, teacher threats or physical control of students, disrespect, bullying, teasing, and sarcasm.
Inclusive class environment	Degree to which the classroom environment is unbiased toward race/ethnicity, religion, gender, sexual orientation, disability, English learner, or other background characteristics, for example as demonstrated in class materials, teacher language, or teacher handling of unacceptable student comments.
Student ease in educational environment	Degree to which students appear comfortable sharing their ideas in class, asking for help, taking risks, or the like; and degree to which there are classroom norms or guidelines that facilitate supportive interactions and feedback.
High expectations for students	Degree to which the teacher imparts high expectations for student learning or shows or describes models of high-quality student work.
Student focus	
Active listening and encouragement of student ideas	Degree to which the teacher actively listens to student comments and responds appropriately (such as summarizing the content of the idea, asking for clarification, asking other students for their thoughts, or expanding on or reinforcing idea); and the teacher encourages students' ideas (for example, by incorporating student responses into the lesson in a meaningful way).
Student autonomy/leadership or productive group work	Degree to which students have choices, leadership opportunities, responsibilities, and delegated roles in class work; all students actively participate in group work; and group work delegates student roles or facilitates meaningful interactions among peers.
Teacher flexibility and tailoring to student needs	Degree to which the teacher tailors or individualizes support to students or demonstrates flexibility with the lesson (such as structure, focus, topic, or timing) to accommodate students' ability levels or understanding.
Family engagement	Degree to which the teacher involves family in student learning, when appropriate.

(continued)

Table C2. Instrument content analysis: Focused-coding scheme with definitions developed through inductive coding *(continued)*

Code	Definition
Classroom management	
Behavior management	Degree to which the teacher states rules and expectations, monitors student behavior, anticipates behavioral problems, or subtly redirects misbehavior; and degree to which students behave well, comply with the teacher's instructions, or otherwise demonstrate understanding of classroom rules and expectations
Time management	Whether the teacher prepares lesson materials in advance and sets up the classroom. Degree to which the teacher knows the lesson, efficiently completes managerial tasks like taking attendance, minimizes time spent on disruptions, and provides ongoing tasks for students. Degree to which the students know what they have to do and how to do it. Degree to which transitions between class activities are short and implemented efficiently.
Active student participation in class activities	
Active student participation in class activities	Degree to which students are actively engaged in class activities most of the time, as demonstrated by students paying attention, raising hands, answering questions, participating in group work, or the like.
Student intellectual engagement with content	
Cognitive challenge	Degree to which the tasks intellectually engage and sufficiently challenge students, without being too difficult or grade-level inappropriate or leaving them completely confused; and degree to which students work to solve challenging problems or complete tasks requiring higher-order thinking.
Student connection questions or alternative ideas	Degree to which students ask questions that seek to explore, draw connections, or identify strategies for solutions; or students offer alternatives to ideas presented by the teacher or other students (such as differentiating their point of view or challenging another's point of view).
Student explanation, prediction, or investigation	Degree to which students explain an approach to solve a problem, a response or answer, their thinking process, or the meaning of an answer; explanations focus on the "why" rather than the "how," use evidence to develop claims, and include self-evaluation or self-reflection; students investigate problems, analyze/interpret information, hypothesize or brainstorm, or work on other open-ended tasks or investigations (that is, problem-based approaches to explore concepts in an in-depth way).
Student perseverance	Degree to which students persevere with their work, even when the task is challenging.
Lesson structure and facilitation	
Clear learning objectives	Degree to which the teacher clearly communicates learning objectives or re-orients students to objectives; and students are aware of the lesson's purpose.
Clear presentation, sequence, or effective pacing	Degree to which the teacher presents information in a clear and organized manner, stays on topic, keeps the lesson moving, or does not allow students to spend excessive time on one task. Whether the lesson includes engagement, learning, and closure stages. Degree to which there is sufficient time for student development of ideas, reflection, and closure.
Variety of strategies, materials, or efforts to engage students	Degree to which the teacher uses a variety of strategies or approaches (such as presentation, questioning strategies, small-group discussion) and types/formats of materials (such as slideshow versus artifacts) to actively engage all students in the lesson; the teacher demonstrates interest in the content and in students' work and ideas; the lesson strategies and materials are appropriate to the learning objectives; students interact with materials; and the teacher effectively integrates support staff to assist with activities.
Effective classroom setup	Degree to which the teacher arranges the classroom and procedures in a way that supports learning (for example, so that all students can participate in large group discussion) or promotes safety (such as the physical setup of a science lab).
Content understanding	
Explicit, in-depth explanation of concepts or procedures	Degree to which the teacher provides explicit explanation of concepts or procedures when explaining them (such as by defining and explaining the conditions under which the concept or procedure is used) and spends several minutes explaining why a procedure works, why an answer is correct, or the like, in a way that conveys meaning.
Rich and meaningful content	Degree to which lesson content is meaningful and focuses on the meaning of facts, procedures, skills, or key practices that are important for students to know.

(continued)

Table C2. Instrument content analysis: Focused-coding scheme with definitions developed through inductive coding *(continued)*

Code	Definition
Connections to real world or prior knowledge and experiences	Degree to which the lesson content is connected to the real world or to students' everyday lives or their prior knowledge or skills related to the concepts being taught (such as what they previously learned).
Use of multiple representations or examples	Degree to which the teacher illustrates a concept, idea, or procedure by providing multiple examples and non-examples or by offering or asking for more than one representation or perspective on the concept (such as two sides of a story in a history lesson or links to another concept or discipline).
Opportunities to practice applying concepts or procedures	Degree to which the teacher provides students the opportunity to practice applying concepts, procedures, or skills, whether independently or supervised by the teacher.
Teacher content understanding and accuracy	Degree to which the teacher possesses deep knowledge of the lesson content, uses appropriate examples and discussion probes, and presents verbal and written information without errors.
Language and discourse	
In-depth, content-driven student discussion	Degree to which students engage in in-depth discussions about the content (such as discussing strategies and providing thoughtful responses to others' ideas).
Use of academic language	Degree to which academic or technical vocabulary is defined, encouraged, and used often.
Discussion facilitation and cumulative exchanges	Degree to which the teacher and students facilitate discussion (such as through open-ended questions, active listening, acknowledgement, pauses, and the like); students assume an active role in conversations; comments in conversation build on each other; and conversation builds knowledge (for example, students and teachers actively listen to each other and add comments that respond to and build on previous comments).
Feedback and assessment	
Back and forth teacher–student feedback exchanges	Degree to which the teacher and students engage in back-and-forth feedback exchanges (such as when the teacher asks follow-up questions to student responses while providing feedback).
Scaffolding feedback	Degree to which the teacher or other students offer hints or other assistance to a student who is unable to produce a correct answer.
Specific feedback	Degree to which the teacher provides specific feedback to a student response, by clarifying or expanding on it (such as by identifying the source of an error, explaining or demonstrating the correct procedure, or clarifying meaning).
Correcting student misconceptions	Degree to which the teacher calls attention to a common misconception before students make errors.
Checking for student understanding	Degree to which the teacher creates opportunities to formally or informally check student understanding of content (such as through formative assessments, by observing group or individual work, in class discussion, or by analyzing records of student work).
Assessment linked to objectives	Degree to which formative and summative assessments are linked to the objectives of the content (where formative assessments are used to gather information to inform improvements, while summative assessments measure success or proficiency, usually at the end of a unit).
Teacher professionalism	
Reflective teacher practice	Degree to which the teacher is reflective of his or her instructional practice and can identify both strengths and weaknesses of a lesson after it is completed.
Collaboration, leadership, and professional learning	Degree to which the teacher actively participates in professional development opportunities to update knowledge and skills, demonstrates intellectual engagement outside the classroom, and collaborates with teacher colleagues to develop teaching practice (such as through a professional learning community), participates in community events, contributes to department decisionmaking, or takes on other leadership roles.
Integrity with colleagues and parents	Degree to which the teacher demonstrates rapport and mutual trust with colleagues and parents.

Source: Authors' inductive coding of the Classroom Assessment Scoring System, Framework for Teaching, Protocol for Language Arts Teaching Observations, Mathematical Quality of Instruction, and UTeach Observational Protocol observation instrument rubrics.

Table C3. Instrument content analysis: Example coded units from observation instruments for each dimension of instructional practice

Dimension	Example text from instrument rubric	Data source
Supportive learning environment	"A safe environment for student risk taking"	FFT
	"-Checks in with students -Anticipates problems -Notices difficulty"	CLASS
Student focus	"-Shows flexibility -Follows students' lead -Encourages student ideas and opinions"	CLASS
Classroom management	"Teacher behaviors…include setting clear behavioral expectations for students and making sure these expectations are met, foreseeing and preparing for inappropriate behavior that may occur during the course of the lesson, consistently and effectively dealing with off-task and inappropriate behavior..."	UTOP
	"The extent to which lesson time is used efficiently; class is on task"	MQI
Active student engagement in class	"Students are eager to participate in the lesson. They raise their hands or call out answers. Most students are engaged in this fashion."	MQI
Student intellectual engagement with content	"Learning tasks require students to engage intellectually, to think; some may involve productive struggle."	FFT
Lesson structure and facilitation	"Purpose of lesson is clear. Its immediacy and relevance, as well as its tie-in to broader purposes, are also explicit."	PLATO
	"… The teacher has chosen and uses appropriate resources to successfully implement the lesson. The evidence gathered should demonstrate that the teacher carefully selected resources that enhance the learning opportunities of the students, while avoiding resources that serve as distractions …"	UTOP
Content understanding	"… The teacher demonstrates deep knowledge and fluidity with the content, as evidenced by the teacher giving detailed and clear explanations, using the big ideas of the content area as a unifying theme, calling attention to applications of the concepts being taught, and fluidly using examples and connections within the subject area."	UTOP
Language and discourse	"Teacher consistently models use of academic language and terms. Students can be heard using terms correctly and flexibly."	PLATO
	"Opportunities for elaborated, conversations between teacher and students and among students are consistent, natural and appropriate to lesson goals. Focus is clear, and stays on track."	PLATO
Feedback and assessment	"Teacher feedback is timely and specific. It explicitly acknowledges what students did well, and addresses problems/incomplete understanding."	PLATO
	"…teacher uses formative assessment techniques to gain awareness of his or her students' progress and understanding."	UTOP
Teacher professionalism	"Collaboration with colleagues for joint planning, and school/district and community initiatives."	FFT

CLASS is Classroom Assessment Scoring System. FFT is Framework for Teaching. PLATO is Protocol for Language Arts Teaching Observations. MQI is Mathematical Quality of Instruction. UTOP is UTeach Observational Protocol.

Source: Authors' deductive coding of the CLASS, FFT, PLATO, MQI, and UTOP observation instrument rubrics.

Table C4. Selected results for teacher value-added scores, summarized across district-specfic models

Model	Mean coefficient	Mean standard error	Percentage significant	*R* squared	Number of teachers
Grades 4–5					
English language arts					
Year 1					
Mean	0.00	0.20	48 percent	0.64	149
Standard deviation	0.01	0.05	39 percentage points	0.06	79
Year 2					
Mean	0.00	0.19	29 percent	0.66	95
Standard deviation	0.01	0.04	33 percentage points	0.07	40
Math					
Year 1					
Mean	0.00	0.20	28 percent	0.67	141
Standard deviation	0.01	0.02	10 percentage points	0.05	74
Year 2					
Mean	-0.01	0.19	44 percent	0.67	88
Standard deviation	0.03	0.02	34 percentage points	0.05	33
Grades 6–8					
English language arts					
Mean	0.00	0.10	19 percent	0.64	131
Standard deviation	0.02	0.02	12 percentage points	0.07	43
Math					
Mean	-0.01	0.09	46 percent	0.66	119
Standard deviation	0.01	0.01	12 percentage points	0.09	36
Grade 9					
English language arts					
Mean	0.00	0.08	8 percent	0.68	63
Standard deviation	0.00	0.00	2 percentage points	0.07	31
Math					
Mean	-0.01	0.08	16 percent	0.53	47
Standard deviation	0.01	0.01	8 percentage points	0.02	8

Note: Sample includes the 117,703 students assigned to the 5,409 teachers included in any of the correlation analyses to assess the relationship between teachers' value-added and observation scores. Reported results summarize estimates of teacher value added (mean coefficient), precision (mean standard error), and model fit (*R*-squared) across the number of teachers included in a given model. Results of two-tailed significance tests at the .05 level are reported as the percentage of significant teacher value-added scores per model.

Source: Authors' analysis of Measures of Effective Teaching data.

Table C5. Supplementary results for the strength of relationship between teachers' overall observation dimension scores and value added to student learning

	Mean across all grades and subjects		Grades 4 5		Grades 6 9	
			English language arts	Math	English language arts	Math
Dimension	Rho	Adjusted rho	Rho	Rho	Rho	Rho
Supportive learning environment	0.10	0.17	0.10	0.12	0.08	0.10
Classroom management	0.15	0.26	0.14	0.16	0.12	0.19
Student intellectual engagement with content	0.11	0.18	0.05	0.11	0.15	0.11
Lesson structure and facilitation	0.11	0.18	0.12	0.13	0.07	0.11
Content understanding	0.06	0.11	0.01	0.09	0.09	0.06
Language and discourse	0.09	0.15	0.11	0.11	0.06	0.07
Feedback and assessment	0.13	0.22	0.19	0.13	0.08	0.11

Note: Sample includes all teachers who taught a class with at least one valid observation instrument score and who taught at least 5 students with a valid state assessment outcome score. Grade 4–5 teachers in district 4 were excluded from these analyses because data on eligibility for the federal school lunch program were missing for all students. Reported results summarize correlations between teachers' value-added scores and their overall dimension score, across the observation instruments with eligible scores for a given dimension (rho). The adjusted correlation (adjusted rho) was calculated as rho divided by the square root of the interyear correlation in value added when estimated separately for year 1 and year 2 (see appendix A for details). Correlations were estimated separately by subject (English language arts or math) and primary and secondary grade levels (grades 4–5 or 6–9). For grades 4–5, correlations were estimated between value-added scores in one year and observation scores in another year for all available year combinations (year 1 value-added scores with year-2 observation scores, and vice versa). The reported results summarize findings across all grades and subjects as the mean of the rho and adjusted rho estimates for each of the four subject-by-grade combinations.

Source: Authors' analysis of Measures of Effective Teaching data.

Table C6. Strength and consistency of the relationship between teachers' value-added scores and observation dimension scores, by instrument

Instrument	Mean rho	Mean adjusted rho	Number of correlations	Percentage of significant correlations
Classroom Assessment Scoring System	0.08	0.13	42	38
Framework for Teaching	0.11	0.19	30	40
Protocol for Language Arts Teaching Observations	0.09	0.15	15	27
Mathematical Quality of Instruction	0.05	0.10	12	17
UTeach Observational Protocol	0.19	0.40	15	47

Note: Sample includes all teachers who taught a class with at least one valid observation instrument score and who taught at least five students with a valid state assessment outcome score. Grade 4–5 teachers in district 4 were excluded from these analyses because data on eligibility for the federal school lunch program were missing for all students. The table reports results from correlations between teachers' value-added scores and their instrument-specific dimension score, with and without adjusting for measurement error (see appendix A for details). See table A5 in appendix A for a list of the available subscores by instrument and dimension. Correlations were estimated separately by subject (English language arts or math) and primary and secondary grade levels (grades 4–5 or 6–9). For grades 4–5, correlations were estimated between value-added scores in one year and observation scores in another year for all available year combinations (year 1 value-added scores with year 2 observation scores, and vice versa). The reported results summarize findings across all available correlations for a given instrument—that is, across subjects (English language arts and math), grade-level groupings (grades 4–5 and grades 6–9), year combinations, and dimension scores. The mean rho is the average of all available correlation coefficients for a given instrument, weighted by the number of teachers in each correlation sample. The mean adjusted rho is the average of all correlation coefficients for a given instrument after adjusting for measurement error, weighted by the number of teachers in each correlation sample. The table also reports the total number of correlation coefficients per instrument and the percentage of coefficients that were statistically significant (two-tailed test at the .05 level).

Source: Authors' analysis of Measures of Effective Teaching data.

Table C7. Complete results for the relationship between teacher observation scores and classroom composition

Student characteristic, subject, and instrument	Dimension	Point estimate	Standard error	*p* value	Sample size
Average baseline English language arts score					
English language arts					
Classroom Assessment Scoring System	Overall	0.13	0.12	.31	457
	Supportive learning environment	0.14	0.13	.28	457
	Classroom management	0.17	0.10	.10	457
	Lesson structure and facilitation	0.07	0.12	.54	457
	Student intellectual engagement with content	0.05	0.12	.65	457
	Language and discourse	0.06	0.13	.65	457
	Feedback and assessment	0.02	0.13	.90	457
	Content understanding	0.02	0.13	.90	457
Framework for Teaching	Overall	0.28**	0.11	<.01	457
	Supportive learning environment	0.23*	0.10	.03	457
	Classroom management	0.12	0.10	.24	457
	Lesson structure and facilitation	0.37**	0.11	<.01	457
	Language and discourse	0.31**	0.12	.01	457
	Feedback and assessment	0.21*	0.09	.02	457
Protocol for Language Arts Teaching Observations	Overall	0.02	0.12	.88	457
	Classroom management	0.03	0.10	.79	457
	Lesson structure and facilitation	–0.07	0.12	.55	457
	Student intellectual engagement with content	0.14	0.11	.23	457
	Language and discourse	0.13	0.12	.31	457
	Content understanding	–0.16	0.11	.17	457
Math					
Classroom Assessment Scoring System	Overall	–0.08	0.14	.58	396
	Supportive learning environment	–0.06	0.16	.72	396
	Classroom management	>–0.01	0.13	>.99	396
	Lesson structure and facilitation	–0.05	0.13	.71	396
	Student intellectual engagement with content	–0.21	0.16	.18	396
	Language and discourse	–0.09	0.16	.59	396
	Feedback and assessment	–0.08	0.17	.65	396
	Content understanding	–0.07	0.15	.65	396
Framework for Teaching	Overall	0.07	0.16	.66	396
	Supportive learning environment	0.01	0.17	.93	396
	Classroom management	0.18	0.14	.20	396
	Lesson structure and facilitation	0.03	0.18	.87	396
	Language and discourse	–0.04	0.19	.82	396
	Feedback and assessment	0.07	0.20	.72	396
Mathematical Quality of Instruction	Overall	0.10	0.18	.57	396
	Lesson structure and facilitation	–0.33*	0.14	.02	355
	Student intellectual engagement with content	0.06	0.21	.79	396
	Feedback and assessment	–0.09	0.17	.59	396
	Content understanding	–0.27	0.18	.13	396

(continued)

Table C7. Complete results for the relationship between teacher observation scores and classroom composition *(continued)*

Student characteristic, subject, and instrument	Dimension	Point estimate	Standard error	*p* value	Sample size
Average baseline math score					
English language arts					
Classroom Assessment Scoring System	Overall	0.12	0.13	.37	457
	Supportive learning environment	0.06	0.13	.65	457
	Classroom management	0.20	0.10	.06	457
	Lesson structure and facilitation	0.07	0.12	.59	457
	Student intellectual engagement with content	0.03	0.13	.83	457
	Language and discourse	<0.01	0.13	.97	457
	Feedback and assessment	–0.03	0.14	.81	457
	Content understanding	<0.01	0.14	.99	457
Framework for Teaching	Overall	0.20	0.11	.09	457
	Supportive learning environment	0.17	0.11	.12	457
	Classroom management	0.09	0.11	.43	457
	Lesson structure and facilitation	0.28*	0.12	.02	457
	Language and discourse	0.17	0.12	.16	457
	Feedback and assessment	0.12	0.10	.27	457
Protocol for Language Arts Teaching Observations	Overall	–0.01	0.12	.91	457
	Classroom management	<0.01	0.10	.99	457
	Lesson structure and facilitation	–0.07	0.12	.54	457
	Student intellectual engagement with content	0.07	0.11	.52	457
	Language and discourse	0.10	0.13	.45	457
	Content understanding	–0.15	0.11	.19	457
Math					
Classroom Assessment Scoring System	Overall	–0.14	0.13	.28	396
	Supportive learning environment	–0.06	0.15	.71	396
	Classroom management	–0.07	0.12	.57	396
	Lesson structure and facilitation	–0.08	0.12	.52	396
	Student intellectual engagement with content	–0.23	0.13	.09	396
	Language and discourse	–0.12	0.15	.42	396
	Feedback and assessment	–0.08	0.16	.62	396
	Content understanding	–0.20	0.14	.17	396
Framework for Teaching	Overall	–0.07	0.16	.64	396
	Supportive learning environment	–0.16	0.15	.29	396
	Classroom management	0.05	0.13	.71	396
	Lesson structure and facilitation	–0.09	0.20	.65	396
	Language and discourse	–0.12	0.18	.52	396
	Feedback and assessment	–0.01	0.21	.97	396
Mathematical Quality of Instruction	Overall	–0.19	0.18	.30	396
	Lesson structure and facilitation	–0.39*	0.16	.02	355
	Student intellectual engagement with content	0.11	0.21	.58	396
	Feedback and assessment	>–0.01	0.18	.98	396
	Content understanding	–0.18	0.16	.26	396

(continued)

Table C7. Complete results for the relationship between teacher observation scores and classroom composition *(continued)*

Student characteristic, subject, and instrument	Dimension	Point estimate	Standard error	*p* value	Sample size
Percentage of students who are racial/ethnic minority students					
English language arts					
Classroom Assessment Scoring System	Overall	−0.50*	0.24	.04	457
	Supportive learning environment	−0.59*	0.24	.02	457
	Classroom management	−0.38*	0.19	.04	457
	Lesson structure and facilitation	−0.31	0.24	.20	457
	Student intellectual engagement with content	−0.42	0.24	.08	457
	Language and discourse	−0.36	0.26	.16	457
	Feedback and assessment	−0.28	0.27	.30	457
	Content understanding	−0.24	0.26	.36	457
Framework for Teaching	Overall	−0.71**	0.23	<.01	457
	Supportive learning environment	−0.74**	0.22	<.01	457
	Classroom management	−0.46*	0.20	<.01	457
	Lesson structure and facilitation	−0.71*	0.24	.02	457
	Language and discourse	−0.57*	0.25	.02	457
	Feedback and assessment	−0.54*	0.21	.01	457
Protocol for Language Arts Teaching Observations	Overall	−0.20	0.22	.37	457
	Classroom management	−0.20	0.19	.28	457
	Lesson structure and facilitation	0.18	0.24	.44	457
	Student intellectual engagement with content	−0.34	0.22	.13	457
	Language and discourse	−0.30	0.24	.22	457
	Content understanding	−0.05	0.22	.83	457
Math					
Classroom Assessment Scoring System	Overall	−0.49	0.37	.18	396
	Supportive learning environment	−0.21	0.39	.59	396
	Classroom management	−0.41	0.28	.15	396
	Lesson structure and facilitation	−0.43	0.35	.21	396
	Student intellectual engagement with content	−0.48	0.50	.34	396
	Language and discourse	−0.44	0.47	.35	396
	Feedback and assessment	−0.08	0.54	.88	396
	Content understanding	−0.56	0.38	.15	396
Framework for Teaching	Overall	−0.44	0.54	.42	396
	Supportive learning environment	−0.50	0.53	.35	396
	Classroom management	−0.41	0.39	.29	396
	Lesson structure and facilitation	−0.35	0.58	.54	396
	Language and discourse	−0.29	0.62	.64	396
	Feedback and assessment	−0.21	0.60	.73	396
Mathematical Quality of Instruction	Overall	0.42	0.60	.49	396
	Lesson structure and facilitation	0.22	0.29	.45	355
	Student intellectual engagement with content	0.30	0.58	.60	396
	Feedback and assessment	0.60	0.48	.22	396
	Content understanding	0.56	0.50	.27	396

(continued)

Table C7. Complete results for the relationship between teacher observation scores and classroom composition *(continued)*

Student characteristic, subject, and instrument	Dimension	Point estimate	Standard error	*p* value	Sample size
Percentage of students who are eligible for the federal school lunch program					
English language arts					
Classroom Assessment Scoring System	Overall	−0.34	0.40	.40	350
	Supportive learning environment	−0.38	0.47	.43	350
	Classroom management	−0.24	0.31	.44	350
	Lesson structure and facilitation	−0.25	0.38	.50	350
	Student intellectual engagement with content	0.08	0.43	.86	350
	Language and discourse	−0.17	0.42	.69	350
	Feedback and assessment	−0.24	0.43	.58	350
	Content understanding	<0.01	0.46	.99	350
Framework for Teaching	Overall	−0.58	0.36	.11	350
	Supportive learning environment	−0.40	0.34	.24	350
	Classroom management	−0.43	0.35	.22	350
	Lesson structure and facilitation	−0.58	0.39	.14	350
	Language and discourse	−0.63	0.39	.11	350
	Feedback and assessment	−0.54	0.34	.12	350
Protocol for Language Arts Teaching Observations	Overall	0.33	0.39	.40	350
	Classroom management	−0.02	0.30	.96	350
	Lesson structure and facilitation	0.82	0.43	.06	350
	Student intellectual engagement with content	−0.14	0.43	.75	350
	Language and discourse	−0.07	0.43	.87	350
	Content understanding	0.76	0.44	.09	350
Math					
Classroom Assessment Scoring System	Overall	−0.27	0.42	.52	313
	Supportive learning environment	−0.08	0.48	.13	313
	Classroom management	−0.54	0.35	.21	313
	Lesson structure and facilitation	−0.55	0.43	.74	313
	Student intellectual engagement with content	−0.16	0.43	.87	313
	Language and discourse	−0.21	0.46	.71	313
	Feedback and assessment	0.13	0.49	.65	313
	Content understanding	−0.13	0.41	.80	313
Framework for Teaching	Overall	−0.90	0.53	.09	313
	Supportive learning environment	−0.80	0.52	.12	313
	Classroom management	−0.78	0.41	.06	313
	Lesson structure and facilitation	−0.82	0.64	.20	313
	Language and discourse	−0.58	0.58	.32	313
	Feedback and assessment	−0.81	0.54	.14	313
Mathematical Quality of Instruction	Overall	0.31	0.77	.69	313
	Lesson structure and facilitation	1.10	0.75	.15	272
	Student intellectual engagement with content	−0.58	0.56	.30	313
	Feedback and assessment	1.02	0.51	.05	313
	Content understanding	0.27	0.59	.65	313

* Significant at $p < .05$; ** significant at $p < .01$.

Note: Sample includes students who were randomly assigned to a classroom in year 2 of the study. Reported results show the point estimates from regressing the standardized scores of a teacher observation instrument (overall or dimension specific), by subject, on the average classroom characteristics for students who were randomly assigned to a classroom. The point estimate is the coefficient on the student characteristic variable. Robust standard errors are reported. See table A5 in appendix A for a list of the available subscores by instrument and dimension.

Source: Authors' analysis of Measures of Effective Teaching data.

Table C8. Strength and consistency of relationship between observation dimension scores and classroom composition, by instrument and composition measure

Student characteristic and instrument	English language arts classrooms (all grades)			Math classrooms (all grades)		
	Mean coefficient	Total number scores	Percentage significant	Mean coefficient	Total number scores	Percentage significant
Average baseline English language arts score						
CLASS	0.07	7	0	–0.08	7	0
FFT	0.25	5	80	0.05	5	0
PLATO[a]	0.01	5	0	na	na	na
MQI[b]	na	na	na	–0.16	4	25
Average baseline math score						
CLASS	0.05	7	0	–0.12	7	0
FFT	0.17	5	20	–0.06	5	0
PLATO[a]	–0.01	5	0	na	na	na
MQI[b]	na	na	na	–0.12	4	25
Percentage of students who are racial/ethnic minority students						
CLASS	–0.37	7	29	–0.37	7	0
FFT	–0.60	5	100	–0.35	5	0
PLATO[a]	–0.14	5	0	na	na	na
MQI[b]	na	na	na	0.42	4	25
Percentage of students who are eligible for the federal school lunch program						
CLASS	–0.17	7	0	–0.22	7	0
FFT	–0.52	5	0	–0.76	5	0
PLATO[a]	0.27	5	0	na	na	na
MQI[b]	na	na	na	0.45	4	25

na is not applicable.

CLASS is Classroom Assessment Scoring System. FFT is Framework for Teaching. PLATO is Protocol for Language Arts Teaching Observations. MQI is Mathematical Quality of Instruction. UTOP is UTeach Observational Protocol.

Note: Sample includes students who were randomly assigned to a classroom in year 2 of the study. The table reports results from regressing the standardized scores of a teacher observation instrument (overall or dimension specific), by subject, on the average classroom characteristics for students who were randomly assigned to a classroom. Results are reported separately for English language arts and math classrooms. The mean coefficient is the average point estimate on the student characteristic variable across all dimension scores available for the instrument. The total number of dimension scores and the percentage of scores that were statistically significant (two-tailed test at the 0.05 level, using robust standard errors) are also reported. See table A5 in appendix A for a list of the available subscores by instrument and dimension.

a. Used to rate instruction only in English language arts classrooms.

b. Used to rate instruction only in math classrooms.

Source: Authors' analysis of Measures of Effective Teaching data.

Notes

The authors would like to acknowledge the contributions of Mathematica Policy Research staff, Jenny Chen and Przemyslaw Nowaczyk, who wrote statistical programs for the quantitative analyses; Daisy Gonzalez, who conducted focused coding for the qualitative analysis; Duncan Chaplin, who reviewed the report; Hanley Chiang, Matt Johnson, Steve Lipscomb, Dana Rotz, Elias Walsh, and Clare Wolfendale, who provided methodological advice; and Felita Buckner, who prepared the report for production.

1. Steinberg and Garrett (2016) use Measures of Effective Teaching data to examine the relationship between classroom characteristics and practice ratings, but they examine only the Framework for Teaching, and their primary analysis relies on within-teacher comparisons across classes rather than random assignment.
2. Instrument-specific observation scores were constructed as described above, from subscores defined by instrument developers. The analysis of dimensions was again limited to those with a dimension score measured by more than one instrument. See appendix A for additional details.
3. Some classrooms were taught by subject matter generalists who were evaluated on both English language arts and math observation instruments.
4. Prior to conducting the analyses, the study team conducted a falsification test that confirmed the general success of random assignment, finding no more significant results than would be expected by chance. The procedure is described in detail in appendix A.
5. The study estimated regressions of Protocol for Language Arts Teaching Observations scores for five dimension scores plus one overall instrument score on each of the four student composition characteristics, for a total of six regressions per classroom composition measure.
6. Although the MET project focused on school years 2009/10 and 2010/11, it collected data on a subset of roughly 350 teachers who volunteered to participate over the following two school years (2011/12 and 2012/13), but those data are not available to outside researchers.
7. This is a teacher fixed-effects approach, which uses only within-teacher variance to identify the relationships between student achievement and student or classroom characteristics. While the MET longitudinal database includes teacher value-added scores, these were estimated by the MET research team using an approach akin to a random-effects approach because it uses both within-teacher and within-school (between teacher) variance to identify these effects (Kane & Staiger, 2012). If certain types of teachers are systematically assigned to certain types of students and classrooms, this approach will produce biased estimates of associations between achievement and student/classroom characteristics. Therefore, the study team preferred the fixed-effects approach; however, identification requires sufficient within-teacher variation in student/classroom characteristics. Thus, this study measured teacher value-added only for teachers with MET data from more than one class section in the dataset. Teachers could contribute data on multiple class sections from one school year, most often applied to teachers in grades 6–9, or class sections from multiple school years, most often applied to teachers in grades 4–5.
8. To make outcome measures from different subjects and assessments comparable, we used standardized test score measures, which the MET researchers scaled by district, year, and grade level so that each has a mean of zero and standard deviation of one. This is particularly important because the state assessments are unique to districts as each was located in a different state (see Kane & Staiger, 2012).

9. Classroom characteristics were excluded from the grade 4–5 models, because these were estimated separately by year and most teachers in these grades taught only one classroom per year.
10. The study estimated the teacher value-added models separately for grade 9, to account for any differences between high school and middle school grades in the production of teacher contributions to student learning. For parsimony, the study then pooled grades 6–9 to estimate correlations between teachers' value-added scores and their observation scores.
11. For example, the CLASS instrument's positive climate subscore is an overall score (scale of 1–7) on relationships, positive affect, positive communications, and respect (Pianta et al., 2012).
12. Of the 10 dimensions of instructional practice that the content analysis identified in the five instrument rubrics (see table 2), the MET data included subscores from more than one instrument for seven dimensions. They would have included dimensions rated by at least two instruments, but all seven dimensions were rated by either one instrument or more than two instruments. Three dimensions had a pre-existing subscore from only one instrument: teacher professionalism (UTOP lesson reflection subscore), student focus (CLASS regard for student perspectives subscore), and active student participation in class activities (CLASS student engagement subscore).
13. Table A5 differs from table 2 in the main text in instrument coverage of specific dimensions due to four key differences between the tables. First, they capture different portions of each instrument. Table A5 reports only data points included in the MET database, which are restricted to instrument components that could be scored based on classroom videos alone and excludes those requiring observation outside the classroom, such as the teacher professionalism component of the FFT rubric. Table 2, in contrast, reports on the entire rubric for each observation instrument. Second, table A5 presents subscores that met this study's analytic sample requirements, so dimension subscores that were available in the MET database for only one instrument were excluded from analyses. For example, the MET database includes a teacher professionalism dimension score only for the UTOP instrument, so this information was not reported in table A5. Third, the tables are based on different versions of the instruments. While table A5 reports coverage of dimensions for the versions used in the MET project, which began in 2009, table 2 reports coverage of dimensions on the most up-to-date instrument rubric that the study team was able to obtain as of spring 2015 (see table A2 for list of versions used). Fourth, the tables capture coverage at different levels of analysis. For each instrument table A5 reports whether the MET database included pre-existing subscores for each dimension, while table 2 reports whether instrument scoring rubrics included language capturing each dimension at the phrase level. Tables 2 and 3 are based on the same data sources.
14. This approach implicitly treats variation in teacher quality over time as error, although some of this variation may be due to changes in teacher practice across years.
15. Each teacher observation score (overall and dimension specific) was standardized using the full sample across both study years with a mean of 0 and a standard deviation of 1.
16. In year 1 of the study, some teachers were observed on the same instrument teaching multiple class sections. In those cases, a teacher's observation score for that instrument was averaged across each observation. In year 2 of the study, there were a few teachers in the randomized sample who taught more than one class section, but the study's analytic sample contained just one class section per teacher.
17. Baseline test scores are from state administered assessments reported as rank-based *z*-scores within district, subject, and grade. Classroom averages were constructed by averaging each student's test score.

References

Borman, G. D., & Kimball, S. M. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *The Elementary School Journal, 106*(1), 3–20. http://eric.ed.gov/?id=EJ725171

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh public schools*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, Regional Educational Laboratory MidAtlantic. http://eric.ed.gov/?id=ED545232

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593–2632.

Danielson, C. (2015). *The framework for teaching: The six clusters*. Princeton, NJ: Danielson Group. Retrieved May 19, 2015, from http://www.danielsongroup.org.

Danielson, C. (2014). *The framework for teaching evaluation instrument*. Princeton, NJ: Danielson Group. Retrieved January 6, 2015, from http://www.danielsongroup.org.

Doherty, K. M., & Jacobs, S. (2013). *State of the states 2013 connect the dots: Using evaluation of teacher effectiveness to inform policy and practice*. Washington DC: National Council on Teacher Quality. http://eric.ed.gov/?id=ED565882

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis, 37*(2), 224–242. http://eric.ed.gov?id=EJ1063556

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2015). *Who believes in me? The effect of student-teacher demographic match on teacher expectations* (Working Paper No. 15–231). Kalamazoo, MI: Upjohn Institute for Employment Research.

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the "Rothstein Falsification Test": Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness,* 88(1), 8–34. http://eric.ed.gov/?id=EJ1049629

Grossman, P., & Greenberg, S. (n.d.) *PLATO: Protocol for Language Arts Teaching Observations*. Stanford, CA: Stanford University Institute for Research on Education Policy & Practice.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445–470. http://eric.ed.gov/?id=EJ1012899

Hill, H. C. (2014). *Mathematical Quality of Instruction (MQI): 4-point version*. Ann Arbor, MI: University of Michigan Learning Mathematics for Teaching Project.

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

Kane, T. J., McCaffrey, D., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Seattle, WA: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project. http://eric.ed.gov/?id=ED540959

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics* 9(3): 1484–1509.

Lunt, M. (2011). *A guide to imputing missing data with Stata: Revision 1.4.* Retrieved July 8, 2015, from http://personalpages.manchester.ac.uk/staff/mark.lunt/mi_guide.pdf.

Lynch, K., Chin, M., & Blazar, D. (2013). *How well do teacher observations of elementary mathematics instruction predict value added? Exploring variability across districts.* Harvard Graduate School of Education working paper. Cambridge, MA.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook.* Thousand Oaks, CA: Sage.

Pianta, R. C., Hamre, B. K., & Mintz, S. L. (2012). *Classroom assessment scoring system upper elementary manual.* Charlottesville, VA: Teachstone.

Rotz, D., Johnson, M., & Gill, B. (2014). *Value-added models for the Pittsburgh public schools, 2012–13 school year.* Report to the Pittsburgh Public Schools. Cambridge, MA: Mathematica Policy Research.

Steinberg, M. P., & Garrett, R. G. (2016). What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis 36*(2): 293–317.

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP middle schools: Impacts on achievement and other outcomes.* Washington, DC: Mathematica Policy Research. http://eric.ed.gov/?id=ED540912

UTeach. (2014). UTeach observation protocol. Austin, TX: University of Texas. Retrieved January 2, 2015, from http://utop.uteach.utexas.edu.

Whitehurst, G. J., Chingos, M. M., & Lindquist, J. M. (2014). *Evaluating teachers with classroom observations, lessons learned in four districts.* Washington, DC: Brown Center on Education Policy at the Brookings Institution. http://eric.ed.gov/?id=ED553815

The Regional Educational Laboratory Program produces 7 types of reports

Making Connections
Studies of correlational relationships

Making an Impact
Studies of cause and effect

What's Happening
Descriptions of policies, programs, implementation status, or data trends

What's Known
Summaries of previous research

Stated Briefly
Summaries of research findings for specific audiences

Applied Research Methods
Research methods for educational settings

Tools
Help for planning, gathering, analyzing, or reporting data or research