Appendix A: Detailed Description of Teacher-level Measures

In this appendix, we provide more detail on the collection of the 22 teacher-level measures included in this study.

Measures of Instruction

We derived observation-based measures of instruction by evaluating video recordings using two established instruments: CLASS (Pianta et al. 2007) and Mathematical Quality of Instruction (MQI; Hill et al. 2008). CLASS is a subject-matter-independent observation tool designed to capture content-general domains of student-teacher interactions, such as classroom organization. MQI, on the other hand, captures mathematics-specific features of instruction, such as teachers' mathematical errors and imprecisions. We followed each instrument's specific protocol to generate scores. For CLASS, one rater scored each 15-minute segment within each recorded lesson on a scale from 1 to 7 for each of the 12 CLASS items. For MQI, two raters scored each 7.5-minute segment on a scale of 1 to 3 for each of the 13 MQI items.

To ensure the standardization of raters' scoring practices, raters (19 for CLASS, 56 for MQI) passed certification exams and attended weekly scoring calibration meetings, each typically an hour long. Raters watched, scored, and wrote rationales on two short video clips before these calibration meetings. We assessed raters' performance on these calibration exercises and provided individualized feedback to raters. Although the primary goal of these sessions was to standardize scoring practices, in a few rare instances we dismissed raters who consistently failed to match master raters' scores. Although inter-rater agreement rates could not be calculated for CLASS since each video is scored by one rater, these processes contributed to the 76% exact-agreement rate on scores for MQI.

Based on prior factor analyses from the same study (Authors 2016), we consolidated the 13 MQI items into two mathematics-specific factors and the 12 CLASS items into two content-general factors (see Appendix Tables A1 and A2 for a description and categorization of each item). To do so, we first averaged scores for each item across all segments and then across raters (for MQI). Then, we averaged the relevant subset of items across all recorded lessons, which produced a teacher-level score for each of these four factors. Finally, we adjusted these four teacher-level scores for reliability to produce an empirical Bayes' shrunken estimate, and then standardized the shrunken scores to have a mean of zero and standard deviation of one. We describe each measure below and report an adjusted intraclass correlation coefficient (ICC) as an estimate of reliability (please see Appendix B for a detailed discussion of the ICC). Although the reliability estimates for the observation-based measures of instructional practices were modest, other studies documented similarly modest reliabilities of contemporary classroom observation instruments (e.g., Bell et al. 2012; Kane and Staiger 2012).

Ambitious Instruction

We measured the extent of ambitious mathematics instruction in teachers' classrooms by using ten items from MQI (e.g., using multiple procedures or solution methods), which evaluate the overall meaning orientation and cognitive demand of a teacher's mathematics instruction. The ICC was 0.69.

Mathematical Errors

To measure the presence of mathematical errors and imprecision in teachers' instruction, we included the remaining three items from the MQI evaluating the extent of the teacher's major

mathematical errors, imprecision in language or notation, and lack of clarity in presentation of mathematical content. The ICC was 0.52.

Classroom Organization

In addition to the two domain-specific classroom observation measures described above, we also generated a content-general measure of classroom organization using three items from CLASS measuring behavior management, productivity, and the negative climate in the classroom. The ICC was 0.65.

Support

We measured a second content-general feature of teachers' classroom practice, which evaluated the emotional and instructional support provided to students. To do so, we included the remaining nine items from CLASS (e.g., student engagement, teacher sensitivity, respect for student perspectives). The ICC was 0.51.

We supplemented the four video-based instructional practices with four survey-based measures of instructional content and its alignment to tested material (see Appendix Table A3 for a description and categorization of each item). All four measures focused on the extent to which classroom content and problem formats matched those on the project-administered assessments. As with the observation-based measures of instruction, we estimated an average score for each teacher across all relevant items, adjusted for reliability, and transformed the score for each measure into a *z*-score.

Algebra Content

We presented teachers with a list of nine algebra topics on the spring survey, and teachers marked which topic(s) were covered in class that year. Topics included using a symbol to stand for an unknown number, using algebraic notation to represent patterns, and identifying a mathematical function from input/output pairs. The ICC was 0.81.

Number and Operations Content

We presented teachers with a list of 16 number and operations topics on the spring survey, and teachers marked which topic(s) were covered in class that year. Topics included understanding place value with decimals, adding and subtracting fractions with like denominators, and converting between decimal form and fraction form. The ICC was 0.85.

Test Prep Activities

We asked teachers five questions regarding the extent to which they engaged in instructional behaviors designed to improve student performance on state standardized tests, such as using released test items or practice test materials. Teachers responded on a 4-point Likert scale from 1 = *never or rarely* to 4 = *daily*. The ICC was 0.81.

Test Prep Instructional Changes

We asked teachers seven questions regarding the extent to which they changed their instructional practices in response to state-imposed testing and accountability systems. For instance, whether they spent more time on mathematics topics that carry more weight on the state test or changed the sequencing of topics so that content more likely to appear on the state test is

covered before the test is administered. Teachers responded on a 5-point Likert scale from 1 = *not at all* to 5 = *very much*. The ICC was 0.87.

Measures of Teacher Personal Characteristics

We collected four measures of teacher personal characteristics using surveys, as described below.

Mathematical Knowledge

To measure teachers' mathematical knowledge, we surveyed teachers using 72 items from the Mathematical Knowledge for Teaching measure (MKT; Hill et al. 2005) and 33 total released items from the mathematics component of a state test for educator licensure (STEL). The MKT assesses teachers' facility in using mathematical knowledge in the context of classroom teaching. For example, items test teachers' ability to select appropriate representations and examples of mathematical concepts such as fractions. The STEL measures subject matter knowledge in the upper elementary and middle grades, such as knowledge of exponents and recognizing patterns. For examples of MKT and STEL items, please see Appendix Figures A1, A2, A3, and A4. To account for the ordinal structure of certain item scores (i.e., testlets), we used a one-parameter graded response model in IRTPRO to generate a single *z*-score for each teacher. The marginal test reliability was 0.85.

Knowledge of Student Performance

Inspired by theories of teacher pedagogical content knowledge (Shulman 1986), we generated a measure of teachers' knowledge of their students' mathematics performance. To do

so, we presented teachers with a subset of items from the project-administered mathematics assessment and then asked what percent of their students would answer the item correctly. In 2010–11, we presented fourth- and fifth-grade teachers with 14 and 15 items, respectively, that appeared on the project-administered assessment. In 2011–12, we presented fourth- and fifth-grade teachers each with eight items. For each of these items, we calculated the absolute difference between teachers' estimated percentage of students correctly answering the item and the actual percentage of students correctly answering the item. To generate a single score for each teacher, we estimated an average score for each teacher across all relevant items, adjusted for reliability, and finally transformed the score for each measure into a *z*-score (which we reversed in this case so that higher scores indicated that the teacher had greater knowledge of student performance). The ICC was 0.89.

Self-efficacy

Based on the work of Tschannen-Moran et al. (1998), we collected 14 items regarding teachers' beliefs about how much they can control classroom behavior, motivate students, and craft good instruction (see Appendix Table A4 for a description and categorization of each item). Teachers responded on a 7-point Likert scale ranging from 1 = *Not at all* to 7 = *A great deal*. For example, one question asked how much they believe they can do to help their students value learning. To generate a single self-efficacy score for each teacher, we calculated the average across these 14 relevant items, adjusted the average for reliability, and finally transformed the score for each measure into a *z*-score. The ICC was 0.73.

Effort

Because research suggests that positive effects of merit pay may operate through teacher effort (Lavy 2009; Muralidharan and Sundararaman 2011), we generated a measure of teacher effort. To do so, we asked teachers to indicate the number of hours per week spent on four non-instructional activities: preparing for class, organizing materials, grading homework, and reviewing the content of lessons (see Appendix Table A4 for a description and categorization of each item). Teachers responded using a 5-point Likert scale ranging from 1 = *None* to 5 = *More than six hours*. To generate a single effort score for each teacher, we calculated the average across these 4 relevant items, adjusted the average for reliability, and finally transformed the score for each measure into a *z*-score. The ICC was 0.82.

Measures of Background

We generated 10 variables describing teachers' educational background and preparation. These measures were derived from 10 survey questions. Of these measures, three were indicator variables (0 = *no*, 1 = *yes*) that captured teachers' course-taking and educational attainment. *Master's degree* indicated any earned master's degree; *math major* indicated an undergraduate or graduate degree in mathematics; and *education bachelor's* indicated a bachelor's degree in education. We also asked teachers to provide both the number of undergraduate- or graduate-level *math courses* and *math content courses* they had taken, using a 4-point scale from 1 = *No Classes* to 4 = *Six+ Classes.*

In addition, we created five indicators of teaching preparation and experience. *Traditional certification* indicated that the teacher currently holds a traditional teaching certification; *alternative certification* indicated that the teacher currently holds an alternative teaching certification; *elementary math certification* indicated that the teacher possesses a

specific certification for teaching elementary mathematics; *4–10 years experience* indicated that the teacher has between 4 and 10 years of teaching experience (including the year surveyed); and *10+ years experience* indicated that the teacher has more than 10 years of teaching experience (including the year surveyed). Unlike the other teacher-level measures, we did not transform these 10 background measures into *z*-scores because these variables had a natural interpretation in their original form. In addition, as these variables were generated from single items, we could not calculate reliability statistics, such as Cronbach's alpha.

Table A1

CLASS Observation Instrument Items

Dimension	Item	Examples of Instruction Measured
CO	Negative Climate	Display of negative affect or punitive control in classroom
CO	Behavior Management	Teacher makes behavioral expectations clear and redirects misbehavior
CO	Productivity	Learning time is maximized and transitions are smooth
S	Student Engagement	Students are focused and participating
S	Positive Climate	Display of positive affect and respect in classroom
S	Teacher Sensitivity	Teacher responds to social/emotional needs of students and demonstrates awareness
S	Respect for Student Perspectives	Teacher connects classroom content to student life and encourages student participation
S	Instructional Learning Formats	Teacher clearly presents materials and promotes involvement and engagement of students
S	Content Understanding	Teacher and students interact in ways that lead to student understanding
S	Analysis and Problem Solving	Teacher facilitates students' use of higher-level thinking skills
S	Quality of Feedback	Feedback from teacher and students promote deeper learning and participation of students
S	Instructional Dialogue	Teacher uses dialogue purposefully to facilitate students' understanding

Note: CO = Classroom Organization; S = Support

Table A2

MQI Observation Instrument Items

Dimension	Item	Examples of Instruction Measured
AI	Linking and Connections	Teacher/students makes links/connections between representations of mathematical ideas
AI	Explanations	Teacher/students gives mathematical meaning to ideas, procedures, steps, or solution methods
AI	Multiple Procedures or Solution Methods	Multiple procedures or solution methods for solving a problem occur or are discussed
AI	Developing Mathematical Generalizations	Teacher/students examine mathematical examples and then make a generalized statement
AI	Mathematical Language	Teacher/students fluently use mathematical language, or such use is supported
AI	Remediation of Student Errors and Difficulty	Student mathematical misconceptions or difficulties are substantially addressed
AI	Responding to Student Mathematical Productions in Instruction	Teacher uses student mathematical productions (e.g., explanations or questions) appropriately
AI	Students Provide Explanations	Students explain why a procedure works or what an answer means
AI	Student Mathematical Questioning and Reasoning	Students provide examples of phenomena or make mathematical conjectures
AI	Enacted Task Cognitive Activation	Level of student engagement with content of the task
E	Major Mathematical Errors	Teacher solves problem or defines term incorrectly
E	Imprecision in Language or Notation	Teacher makes errors in notation, mathematical language, or general language
E	Lack of Clarity in Presentation of Mathematical Content	Teach utterance cannot be understood or launch of task is unclear

Note: AI = Ambitious Instruction; E = Errors

Table A3

Instructional Survey Measures

Dimension	Item
Algebra Content	Please check the box next to the content areas you have covered this year. The meaning of the equals sign as balancing two quantities
Algebra Content	Please check the box next to the content areas you have covered this year. Using algebraic expressions to represent situations (e.g., in a word problem)
Algebra Content	Please check the box next to the content areas you have covered this year. Recognizing and continuing repeating patterns/sequences, or predicting subsequent terms
Algebra Content	Please check the box next to the content areas you have covered this year. Using algebraic notation to represent patterns
Algebra Content	Please check the box next to the content areas you have covered this year. The use of a symbol (e.g., shape or letter) to stand for an unknown number
Algebra Content	Please check the box next to the content areas you have covered this year. Determining a general rule (function) from a series of input/output pairs (e.g., in a table or function machine)
Algebra Content	Please check the box next to the content areas you have covered this year. Interpreting and solving word problems/situations
Algebra Content	Please check the box next to the content areas you have covered this year. Undoing/inverse operations
Algebra Content	Please check the box next to the content areas you have covered this year. Interpreting graphs
Number and Operations	Please check the box next to the content areas you have covered this year. Understanding place value with whole numbers
Number and Operations	Please check the box next to the content areas you have covered this year. Understanding place value with decimals
Number and Operations	Please check the box next to the content areas you have covered this year. Associative, commutative, and/or distributive properties
Number and Operations	Please check the box next to the content areas you have covered this year. Why standard algorithms work (e.g., multi-digit multiplication, long division)

Number and Operations	Please check the box next to the content areas you have covered this year. Non-standard algorithms for basic operations (e.g., partial product method for multi-digit multiplication)
Number and Operations	Please check the box next to the content areas you have covered this year. Multiple representations of decimals (e.g., on a number line, with shaded region` of a figure)
Number and Operations	Please check the box next to the content areas you have covered this year. Comparing or ordering decimals
Number and Operations	Please check the box next to the content areas you have covered this year. Operations with decimals
Number and Operations	Please check the box next to the content areas you have covered this year. Representing fractions graphically (e.g., on a number line, with shaded regions of a figure)
Number and Operations	Please check the box next to the content areas you have covered this year. Meaning of fractions
Number and Operations	Please check the box next to the content areas you have covered this year. Comparing or ordering fractions
Number and Operations	Please check the box next to the content areas you have covered this year. Adding and subtracting fractions with like denominators
Number and Operations	Please check the box next to the content areas you have covered this year. Adding and subtracting fractions with unlike denominators
Number and Operations	Please check the box next to the content areas you have covered this year. Multiplying and dividing fractions
Number and Operations	Please check the box next to the content areas you have covered this year. Converting between decimal form and fraction form
Number and Operations	Please check the box next to the content areas you have covered this year. Comparing the values of decimals and fractions to each other
Test Prep Activities	In a typical week of teaching mathematics, how frequently do you engage in the following activities? : Use test items or practice test materials in preparation for the [state test]
Test Prep Activities	In a typical week of teaching mathematics, how frequently do you engage in the following activities? : Incorporate formats similar to those on the [state test] (such as styles of graphs or key phrases) into my instruction
Test Prep Activities	In a typical week of teaching mathematics, how frequently do you engage in the following activities? : Set aside part of class time to review concepts or skills found on the [state test] (e.g., use a problem of the day)

Test Prep Activities	In a typical week of teaching mathematics, how frequently do you engage in the following activities? : Focus on supporting students who are expected to score just below a given performance level on the [state test]
Test Prep Activities	In a typical week of teaching mathematics, how frequently do you engage in the following activities? : Teach specific test-taking strategies, like process of elimination or plugging in answers
Test Prep Instructional Changes	To what extent does preparing for the [state test] result in the following changes in your instruction? : Spending less time on mathematical topics that are rarely or never tested
Test Prep Instructional Changes	To what extent does preparing for the [state test] result in the following changes in your instruction? : Spending more time on mathematical topics that carry more weight on the test
Test Prep Instructional Changes	To what extent does preparing for the [state test] result in the following changes in your instruction? : Giving students less time to discuss mathematical concepts in depth
Test Prep Instructional Changes	To what extent does preparing for the [state test] result in the following changes in your instruction? : Limiting special projects or hands-on activities related to mathematics
Test Prep Instructional Changes	To what extent does preparing for the [state test] result in the following changes in your instruction? : Using fewer demanding mathematics problems (e.g., "extension" problems) to challenge advanced students
Test Prep Instructional Changes	To what extent does preparing for the [state test] result in the following changes in your instruction? : Exploring mathematical concepts in less depth
Test Prep Instructional Changes	To what extent does preparing for the [state test] result in the following changes in your instruction? : Sequencing mathematical topics so that content usually on the test is covered before the test is administered

Table A4

Personal Characteristic Measures

Dimension	Item
Self-efficacy	To what extent do you agree or disagree with the following statements about your mathematics class? : I can get through to even the most difficult students
Self-efficacy	To what extent do you agree or disagree with the following statements about your mathematics class? : I can craft good questions for my students.
Self-efficacy	To what extent do you agree or disagree with the following statements about your mathematics class? : I can provide an alternative explanation or example when students are confused
Self-efficacy	To what extent do you agree or disagree with the following statements about your mathematics class? : I can use a variety of assessment strategies to help students learn.
Self-efficacy	To what extent do you agree or disagree with the following statements about your mathematics class? : I can implement alternative teaching strategies in my classroom.
Self-efficacy	Please answer these questions based on your current mathematics teaching assignment. : How much can you do to control disruptive behavior in the classroom?
Self-efficacy	Please answer these questions based on your current mathematics teaching assignment. : How much can you do to motivate students who show low interest in school work?
Self-efficacy	Please answer these questions based on your current mathematics teaching assignment. : How much can you do to calm a student who is disruptive or noisy?
Self-efficacy	Please answer these questions based on your current mathematics teaching assignment. : How much can you do to help your students value learning?
Self-efficacy	Please answer these questions based on your current mathematics teaching assignment. : How much can you do to get children to follow classroom rules?
Self-efficacy	Please answer these questions based on your current mathematics teaching assignment. : How much can you do to get students to believe they can do well in school work?
Self-efficacy	Please answer these questions based on your current mathematics teaching assignment. : How well can you establish a classroom management system with each group of students?
Self-efficacy	Please answer these questions based on your current mathematics teaching assignment. : To what extent can you use a variety of assessment materials?

Self-efficacy	Please answer these questions based on your current mathematics teaching assignment. : How much can you assist families in helping their children do well in school?
Effort	In a typical week, how much time do you devote to the following activities? : Grading mathematics assignments
Effort	In a typical week, how much time do you devote to the following activities? : Gathering and organizing mathematics lesson material (e.g., locating and copying supplemental material, preparing manipulatives)
Effort	In a typical week, how much time do you devote to the following activities? : Reviewing the content of specific mathematics lessons (e.g., reading the teacher manual, seeking additional information about the content)
Effort	In a typical week, how much time do you devote to the following activities? : Preparing for a mathematics lesson by trying out explanations, or working through examples of problems

The expression $(5^{-8} \cdot 7^{-9})$ is equal to which of the following? (Circle ONE answer.)

a) $\frac{1}{5(35)^8}$

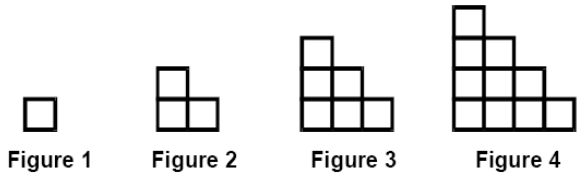
b) $\frac{1}{7(35)^8}$

c) $\frac{5}{(35)^8}$

d) $\frac{7}{(35)^8}$

Figure A1. A sample mathematics state test for educator licensure item from the fall teacher survey testing teachers' mathematical content knowledge of exponents.

Use the diagram below to answer the question that follows.



If the pattern continues, how many more small squares are in figure 100 than are in figure 99? (Circle ONE answer.)

a) 98
b) 99
c) 100
d) 101

Figure A2. A sample mathematics state test for educator licensure item from the fall teacher survey testing teachers' mathematical content knowledge of patterns.

Mr. Foster's class is learning to compare and order fractions. While his students know how to compare fractions using common denominators, Mr. Foster also wants them to develop a variety of other intuitive methods.

Which of the following lists of fractions would be best for helping students learn to develop several different strategies for comparing fractions? (Circle ONE answer.)

a) $\frac{1}{4} \quad \frac{1}{20} \quad \frac{1}{19} \quad \frac{1}{2} \quad \frac{1}{10}$

b) $\frac{4}{13} \quad \frac{3}{11} \quad \frac{6}{20} \quad \frac{1}{3} \quad \frac{2}{5}$

c) $\frac{5}{6} \quad \frac{3}{8} \quad \frac{2}{3} \quad \frac{3}{7} \quad \frac{1}{12}$

d) Any of these would work equally well for this purpose.

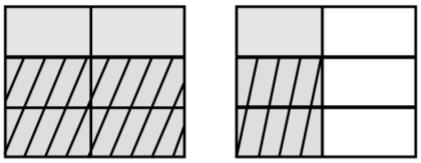
Figure A3. A sample Mathematical Knowledge for Teaching item from the fall teacher survey testing teachers' knowledge of appropriate examples or tasks.

At a professional development workshop, teachers were learning about different ways to represent multiplication of fractions problems. The leader also helped them to become aware of examples that do not represent multiplication of fractions appropriately.

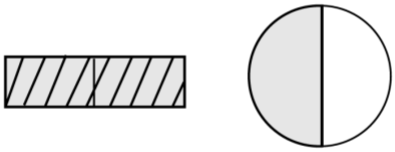
Which model below cannot be used to show that $1\frac{1}{2} \times \frac{2}{3} = 1$? (Mark ONE answer.)

A)

B)



C)



D)

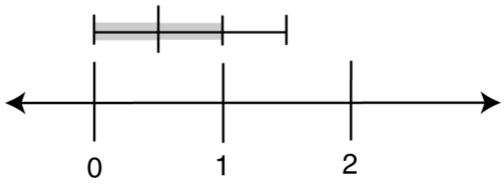


Figure A4. A sample Mathematical Knowledge for Teaching item from the fall teacher survey testing teachers' mathematical knowledge of appropriate representations.

Appendix B: Estimating Intraclass Correlation Coefficients (ICC)

To provide reliability statistics for some of the teacher measures used in our analyses, we estimated ICCs, which provide an estimate of the amount of variance in teacher measure scores attributable to the teacher and not to other identified construct irrelevant sources of variation. The ICC is similar to Cronbach's alpha—another commonly used reliability statistic—as both statistics describe the proportion of total score variance that is due to true score variance. We prefer ICCs to describe teacher measure reliability because it is more flexible; for example, for our teacher scores measured from surveys, we can account for the fact that some teachers answered the same survey prompts over two instances of survey administration in our reliability estimates (see below for more detail).

For teacher instruction scores collected through observation (i.e., scores for the MQI and CLASS measures), the main identified source of construct irrelevant variation is differences between lessons. Thus, to recover the ICC statistics for these scores, we estimate the following equation:

$$p_{j,k} = \beta_0 + \mu_k + \epsilon_{j,k}$$

The outcome, $p_{j,k}$, represents teacher *k*'s MQI/CLASS score for lesson *j*. Parameter μ_k represents teacher *k*'s shrunken random effect on $p_{j,k}$. We do not include a random effect for lesson is included in this model, as lessons are nested within teachers. The ICC can then be calculated for these teacher-scores using the following equation:

$$\text{ICC} = \frac{\text{Var}(\mu_k)}{\text{Var}(\mu_k) + \left(\frac{\text{Var}(\varepsilon_{j,k})}{\bar{n_j}}\right)}$$

We divide the residual variance by the modal number of lessons across sample teachers used to generate scores ($\bar{n_j}$), as this adjustment provides the amount of teacher-level variance in scores

across lessons as opposed to in a single lesson (note that when data are balanced, this adjusted-ICC approximates Cronbach's alpha).

For teacher scores for measures generated from survey responses, the primary source of construct irrelevant variation is differences between survey prompts. Thus, we estimate slightly different model to recover ICCs for teacher measures generated from survey responses:

$$p_{j,k,t} = \beta_0 + \mu_k + \nu_{j,t} + \epsilon_{j,k,t}$$

The outcome, $p_{j,k,t}$, represents teacher *k*'s survey response (e.g., for the self-efficacy or effort items) for survey prompt *j* in year *t*. Parameter μ_k represents teacher *k*'s shrunken random effect on $p_{j,k,t}$. Parameter $\nu_{j,t}$ represents the random effect for survey prompt *j* in year *t*; we include this parameter because teachers and survey prompts are crossed. The ICC can then be calculated for these teacher-scores using the following equation:

$$\text{ICC} = \frac{\text{Var}(\mu_k)}{\text{Var}(\mu_k) + \left(\frac{\text{Var}(\varepsilon_{j,k,t})}{\bar{n_j}}\right)}$$

We divide the residual variance by the modal number of unique survey prompts across teachers used to generate scores ($\bar{n_j}$), as this adjustment provides the amount of teacher-level variance in scores across prompts as opposed to in a single prompt.

Additional Appendix References

Tschannen-Moran, Megan, Anita Woolfolk Hoy, and Wayne K. Hoy. 1998. "Teacher Efficacy: Its Meaning and Measure." *Review of Educational Research* 68(2): 202-248.