American Educational Research Journal
October 2018, Vol. 55, No. 5, pp. 1076–1112
DOI: 10.3102/0002831218769614
Article reuse guidelines: sagepub.com/journals-permissions

Connections Between Teachers' Knowledge of Students, Instruction, and Achievement Outcomes

Heather C. Hill
Mark Chin
Harvard Graduate School of Education

Both scholars and professional standards identify teachers' knowledge of students as important to promoting effective instruction and student learning. Prior research investigates two such types of knowledge: knowledge of student thinking and teacher accuracy in predicting student performance on cognitive assessments. However, the field presents weak evidence regarding whether these constructs are amenable to accurate measurement and whether such knowledge relates to instruction and student outcomes. Without this evidence, it is difficult to assess the relevance of this form of teacher knowledge. In this article, evidence from 284 teachers suggests that accuracy can be adequately measured and relates to instruction and student outcomes. Knowledge of student misconceptions proved more difficult to measure, yet still predicted student outcomes in one model.

Keywords: achievement, instructional practices, mathematics education, teacher knowledge

Knowledge of students ranks high among the teacher capabilities identified by both professional standards documents (Council of Chief School Officers, 2011; National Board for Professional Teaching Standards, 1989) and scholars (Cohen, Raudenbush, & Ball, 2003; Shulman, 1986, 1987) as important to good teaching. Such knowledge is thought to enable a variety of effective classroom strategies, including adjusting the pacing of instruction based on student need (Clark & Peterson, 1986), forming appropriate

Heather C. Hill is the Jerome T. Murphy Professor of Education at the Harvard Graduate School of Education, 6 Appian Way #445, Cambridge, MA 02138; e-mail: *heather_hill@gse.harvard.edu*. She specializes in research on teachers, teacher policy, and mathematics instruction.

Mark Chin is a PhD candidate in Education Policy and Program Evaluation at Harvard University. His research interest centers on the experiences and outcomes of students of color and first- and second-generation students in U.S. K–12 educational contexts.

instructional groups (Shavelson & Borko, 1979), facilely assessing student understanding and misunderstanding in the moment (Ball, Thames, & Phelps, 2008; Johnson & Larsen, 2012), designing instruction to address common student misconceptions (Stein, Engle, Smith, & Hughes, 2008), and designing tasks and questions to further student understanding (An, Kulm, & Wu, 2004; Peterson, Carpenter, & Fennema, 1989). Several broad-scale interventions and professional development efforts are based on the idea that improving teachers' knowledge of students will improve these aspects of instruction and thus student outcomes, and evidence shows that some of these approaches have worked (Bell, Wilson, Higgins, & McCoach, 2010; Black & Wiliam, 1998; Carpenter, Fennema, Peterson, Chiang, & Loef, 1989).

Despite this considerable consensus on the importance of teachers' knowledge of students, research in this domain lacks information on two critical issues. First, the field presents incomplete evidence regarding how well teachers' capacities in this domain can be measured and how such teacher capacities relate to other forms of teacher knowledge, such as subject matter knowledge. Without this evidence, it is difficult to ascertain whether teachers' knowledge of students is a separate, identifiable construct amenable to accurate measurement. Second, the field presents weak and inconsistent evidence regarding whether teachers' knowledge of students relates to teachers' classroom work with students and to student outcomes. Without evidence addressing this second issue, it is difficult to assess claims that teachers' knowledge of students constitutes a key capability for effective teaching. This is true despite nearly three decades of interventions and policies designed, respectively, to improve teachers' knowledge of students and to highlight the role of such teacher knowledge in student learning.

In this paper, we provide evidence on these two critical issues in one subject, mathematics. Following an older literature in educational psychology (Carpenter, Fennema, Peterson, & Carey, 1988; Helmke & Schrader, 1987; Hoge & Coladarci, 1989; Scates & Gage, 1958), we measure teacher *accuracy* in predicting student performance. Following Sadler, Sonnert, Coyle, Cook-Smith, and Miller (2013), we also measure teachers' *knowledge of student misconceptions*. We ask:

1. Do scores on these measures remain stable over time and differentiate among teachers?
2. Do teacher scores on these measures relate to one another and to similar measures of teacher knowledge?
3. Do teacher scores on these measures predict theoretically related measures of instructional quality?
4. Do teacher scores on these measures predict student outcomes?

Following our review of prior research, we discuss the data and analyses that allow us to explore these questions.

Prior Research

Teacher educators have long conceptualized teacher knowledge as a multifaceted construct. Shulman and colleagues' classic formulation of teacher knowledge includes several distinct categories, including content knowledge, general pedagogical knowledge, pedagogical content knowledge, and knowledge of learners and their characteristics, among other topics (Shulman, 1987; Wilson, Shulman, & Richert, 1987). Others scholars followed by elaborating and extending this list (Ball et al., 2008; Rowland, Huckstep, & Thwaites, 2005). Notably, most theories include space for what we here refer to as teachers' *knowledge of students*. For instance, Shulman's (1986) "knowledge of learners" category encompasses "the conceptions and preconceptions that students of different ages and backgrounds bring with them to the learning of those most frequently taught topics and lessons" (p. 9). Within his definition of pedagogical content knowledge, Shulman also counts teacher knowledge of student misconceptions and teacher knowledge of easy and difficult topics for students. Ball et al. (2008) list a more extensive set of elements within what they call "knowledge of content and students" (KCS), including student conceptions and misconceptions around specific content, student interests, and likely student reactions to particular instructional tasks.

In the years since Shulman and colleagues' original formulation, scholars have elaborated how perception, knowledge, and action in this domain intersect. Theories of professional noticing, for instance, define one aspect of teacher expertise as skill in attending to students' strategies, interpreting students' understandings, and responding with appropriate instructional moves (Jacobs, Lamb, & Phillip, 2010; Sherin, Jacobs, & Phillip, 2011). Others have investigated the extent to which teachers watching classroom video focus on mathematical content, teacher moves, student understanding, or other issues (e.g., classroom management; Star & Strickland, 2008). And still others have examined how teachers who both perceive and understand student thinking plan their responses in ways that meet students' needs (Barnhart & van Es, 2015; Jacobs, Lamb, Phillip, & Schappelle, 2011). Figure 1 depicts a simple schema for how teacher knowledge and noticing mutually reinforce one another, and the ways in which teachers may adjust their practice as a result of both. Although we cannot empirically test all components of this schema, it is a useful heuristic for conveying how teacher knowledge, perception, and action may interact.

For this article, we focus in particular on two aspects of teachers' knowledge of students: teachers' knowledge of student thinking and teachers' judgment accuracy. In order to establish knowledge of students as a key teacher characteristic, we argue that scholars must produce three types of evidence: that such knowledge constitutes a stable trait on which teachers differ meaningfully; that this knowledge relates to similar forms of teacher

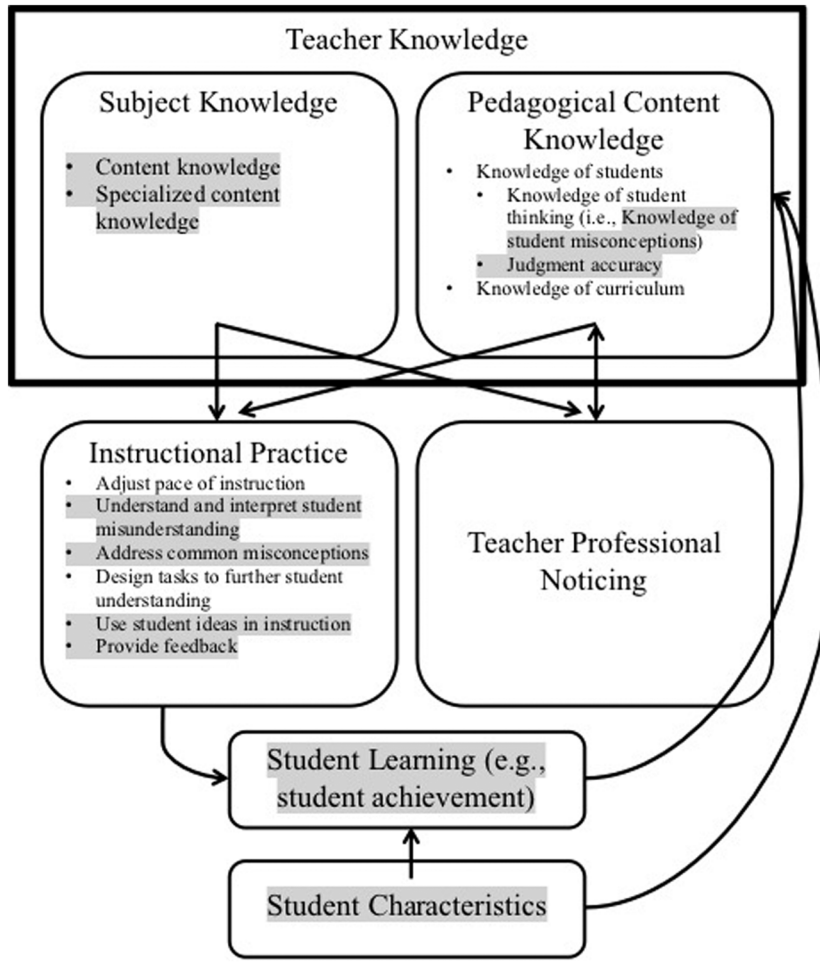


Figure 1. **Schema of relationships that link teacher knowledge to student learning. Measures investigated by our study are highlighted in gray.**

knowledge yet independently predicts instructional quality in expected ways; and that this knowledge predicts student learning, again independently of related forms of teacher knowledge. Lacking such rigorous and conjoined evidence, the field has little assurance that teachers' knowledge of students, as theorized by Shulman and others, will help teachers more facilely execute both in-the-moment and long-term instructional decision-making, as hypothesized above, and that knowledge of students is worth

developing in teacher preparation and assessing at licensure (Shulman, 1986).[1] Below, we report whether and how existing studies provide evidence on teachers' knowledge of students, paying particular attention to measurement strategies, both because they demonstrate how prior work has operationalized teachers' knowledge of students and also because these prior measurement strategies inform our own.

Knowledge of Student Thinking

Some scholars have focused attention on what teachers know regarding student thinking—that is, frequently used student strategies, typical developmental pathways, and common misconceptions. As this definition suggests, teacher knowledge in this domain should mirror empirical findings regarding children's typical mathematical development, and is thus presumed to be a stable teacher-level trait that is transferrable across different classes of students. Conceptualizations of this domain and measurement strategies vary. Some focus heavily on students' developmental pathways and utilize open-ended prompts. For example, Carpenter and colleagues (1988) showed teachers videos of three students solving problems using well-recognized and developmentally appropriate processes (e.g., Carpenter, Moser, & Romberg, 1982) and then asked those teachers to predict how the three students would solve similar problems, crediting teachers for correct predictions. Carpenter et al. (1988) also measured teachers' knowledge of the likely computation strategies deployed by six of their own students, crediting teachers when their predictions matched those students' actual strategies. Bell et al. (2010) focused more on misconceptions, presenting teachers with written student work and asking them to identify and explain student errors, comment on different solution strategies for a single problem, and describe what students might have been thinking as they answered. Teachers received credit on a scoring rubric for more sophisticated responses. Krauss et al. (2008) similarly presented classroom situations and asked teachers to detect, analyze, and predict student misconceptions; this subscale helps compose a larger pedagogical content knowledge scale. Krauss and colleagues (2008) reported no reliability for the student misconception subscale, likely due to the small number of items; Carpenter and colleagues either reported no reliability (Carpenter et al., 1988) or reported that teachers' scores had variable reliability ($\alpha = \{.47; .57; .86\}$; Carpenter et al., 1989). By contrast, Bell et al. (2010) reported acceptable estimates of internal consistency: .76 and .79 on the pretest and posttest, respectively.

Other authors have attempted to measure knowledge of student thinking via multiple-choice items, which allow for efficient measurement at scale. For example, H. C. Hill, Ball, and Schilling (2008) used the same ideas about KCS described in Ball et al. (2008) to design items focused on likely student misconceptions, the source of those misconceptions, problems

that students would find easy or difficult to solve, and common student problem-solving strategies. Using data on more than 1,500 teachers, the authors conducted factor analyses for the purpose of construct validation and to examine convergent and discriminant validity. The factor analysis indicated that teacher performance on the KCS items related to both a mathematical knowledge factor as well as to a specific KCS factor; reliabilities were modest (.58 to .69). Teachers' growth from pretest to posttest correlated to their reports of learning about KCS-related topics in professional development, but not to their reports of learning subject matter knowledge, suggesting convergent and discriminant validity. However, with low score reliability and a strong ceiling effect, these items were not pursued further. Sadler et al. (2013) measured knowledge of students' science misconceptions by asking teachers to identify the most common incorrect student response for each of 20 multiple-choice items. However, the authors did not model knowledge of student misconceptions as a single teacher-level latent trait, instead adopting an item-specific approach described in more detail below.

Despite this healthy interest in measuring teachers' knowledge of student thinking, few studies in the field have related direct measures of such knowledge to similar knowledge constructs, to instruction, and to student outcomes. With regard to similar knowledge constructs, Krauss et al. (2008) demonstrated that teachers' pedagogical content knowledge was related to yet distinguishable from teachers' content knowledge; the correlation between the student-specific pedagogical content knowledge subscales with the content knowledge subscales ranged from .39 to .48. Carpenter and colleagues (1988) found only very modest correlations between different aspects of teacher knowledge of students. With regard to student outcomes, Carpenter and colleagues found no relationship between their two measures of teachers' knowledge of students and actual student performance on tests of computation and solving word problems. This research team did, however, find that teachers with more knowledge of student strategies more often used question-based instructional techniques and listened to student responses (Peterson et al., 1989). Finally, Sadler et al. (2013) compared teacher and student performance on specific items to find that high-achieving students of teachers who possessed both subject matter knowledge and knowledge of student misconceptions posted stronger gains than high-achieving students who had teachers with subject matter knowledge only. There was no such effect for low-achieving students and no main effect of teacher performance on student performance. The authors also graphically demonstrated a strong relationship between science subject matter knowledge and knowledge of misconceptions, but do not report the strength of the correlation.

This article follows on this work by focusing on teachers' knowledge of student misconceptions. We made the choice to narrow to this topic based on Shulman and colleagues' original emphasis on this idea as a key element of teacher knowledge, on our curiosity about whether Sadler et al.'s science

results generalized to mathematics, and on evidence from the field of mathematics education and educational psychology that improving teacher noticing of and actions based on students' misconceptions can improve student outcomes (e.g., Borasi, 1994; Heller, 2010; Jacobs, Franke, Carpenter, Levi, & Battey, 2007). We also chose to capture this construct using multiple-choice items, so that our work linking knowledge of student misconceptions to instruction and student outcomes could proceed at scale. We modify Sadler and colleagues' approach, however, by conceptualizing and measuring teachers' knowledge of student misconceptions not as a set of discrete items, but instead as a unified teacher-level construct, as theorized by Shulman, Ball, and others. We also evaluate the independent effect of teachers' knowledge of students on instruction and student outcomes by including rigorous controls for both student- and classroom-level characteristics and for teacher content knowledge.

Teacher Judgment Accuracy

Concurrent with this substantial interest in knowledge of student thinking, a separate line of work arose from educational psychologists' interest in how teachers' knowledge of their students' performance on specific tested content might support their judgments during instructional decision-making. Here, the conceptualization of teacher knowledge differed from that described above, in that it draws upon several sources of information: knowledge of the particular student(s) in question, knowledge of the content, and knowledge of the testing situation (e.g., high- versus low-stakes). This knowledge is thus contingent (i.e., based on the interaction of several elements), constructed in the moment in order to support instructional decisions such as reteaching material or regrouping students for review. However, there may be important between-teacher differences in the propensity to notice and construct knowledge about student performance, a hypothesis that we test here by drawing on the judgment accuracy literature.

Studies in this tradition utilized a variety of measurement techniques. For instance, Coladarci (1986) asked teachers to anticipate each student's performance on selected achievement test items. The author then differenced the teacher's prediction and student's actual score for each student–item combination before averaging that difference score to the teacher level. Other versions of these metrics asked teachers to predict class rankings or the total scores of each of their students on achievement tests. Analysts typically then correlated teacher estimates of each student's performance with actual performance, with a median correlation of .66 (Hoge & Coladarci, 1989), a finding that suggests that teachers, on average, are relatively accurate in their knowledge of students. Interestingly, however, significant variability across teachers in the accuracy of predictions appeared to exist (Coladarci, 1986; Hoge, 1983; Hoge & Coladarci, 1989; Martínez, Stetcher, & Borko,

2009), leading to a large volume of efforts to identify explanations for such teacher differences (for a review, see Südkamp, Kaiser, & Möller, 2012).

Two studies used this cross-teacher variability in accuracy to predict student performance. Helmke and Schrader (1987) compared 32 fourth- and fifth-grade teachers' predictive accuracy to their students' outcomes. As students completed a mathematics test, teachers estimated, for each student, how many problems on that test the student would solve correctly. The authors then correlated teacher accuracy with student reports that the teacher was aware of their performance levels, suggesting concurrent validity. Teacher accuracy also marginally significantly predicted students' mathematics test scores when interacted with teachers' use of structuring cues and individual support ($p \leq .10$); however, the main effect of teacher knowledge was not significant. Further, this study did not report evidence regarding the reliability of teacher scores or stability of the underlying trait.

Carpenter et al. (1988) constructed a similar measure of judgment accuracy in which teachers predicted whether each of six randomly selected students from their class would successfully solve each of six different addition/subtraction word problems. The authors credited teachers for matches between predictions and actual student outcomes, and then compared these accuracy scores to classroom-aggregated student scores, finding a correlation of roughly .3. However, the models contained neither controls for prior student achievement nor related teacher traits, such as mathematical knowledge, which might independently influence student outcomes, rendering these results open to critiques of selection and omitted variable bias.

Our Study

In sum, prior research in both teacher education and educational psychology has laid the groundwork for our investigation, suggesting facets of teachers' knowledge of students that are amenable to measurement and providing suggestive evidence regarding the importance of such knowledge. However, the reliability of teacher scores on measures in this domain is either not reported or is variable, suggesting that the field lacks comprehensive evidence on the extent to which these constructs can be adequately measured. Similarly, authors seldom report correlations between scores on distinct facets of knowledge of students, or between knowledge of students and related constructs such as teacher content knowledge. Finally, evidence regarding these constructs' relationship to instruction and student outcomes is mixed, with the strongest evidence for the latter appearing in models that do not control for prior student outcomes or similar forms of teacher knowledge (Carpenter et al., 1988), and null or highly contingent results appearing elsewhere (Carpenter et al., 1988; Helmke & Schrader, 1987; Sadler et al., 2013). Weak evidence on score reliability and mixed evidence regarding impact on student outcomes casts doubt on claims that this form of

knowledge underlies effective teaching practice and raises questions about programs and policies aimed to improve such knowledge. This article seeks to address these issues.

Data and Methods

Sample and Setting

We used data collected by the National Center for Teacher Effectiveness, a project that developed and validated mathematics-related measures of teacher effectiveness in fourth- and fifth-grade classrooms. The study invited 583 teachers in four large urban East Coast public school districts to participate; ultimately, 328 of these teachers were eligible for and agreed to participate in the study over a 3-year period (the 2010–2011 academic year through the 2012–2013 academic year). While some teachers participated for all 3 years (n = 130), others participated for only 2 years (n = 79) or only 1 year. In some cases, teachers switched grades between years (e.g., fourth to fifth). Our analytic sample comprised a subset of these teachers and students.

Our primary analysis exploring the relationship between teachers' knowledge of students and student test scores included 284 teachers and 9,636 of their students. This sample excluded teachers working in atypical classrooms, specifically those with more than 50% special education students, those with more than 50% of students missing baseline test scores, and those with fewer than five students included in student test score models. Sample teachers were largely White (65%), female (83%), and entered the teaching profession through traditional teacher education programs (85%). The average teacher had 10 years of teaching experience, and 76% possessed a master's degree. Students were primarily non-White (40% Black, 23% Hispanic) and eligible for free or subsidized lunch (63%). Approximately 10% of these students were special education students, and double that were English language learners (20%).

Data

Project staff collected data from five sources: self-administered teacher questionnaires, digitally recorded mathematics lessons, student questionnaires, a project-developed student mathematics test, and district administrative data. We describe each below.

Teachers responded to questionnaires in the fall and spring semesters of each school year. Consistent with theories about the structure and content of teachers' knowledge (Ball et al., 2008; Shulman, 1986; Wilson et al., 1987) the fall questionnaire measured grade-appropriate content and specialized content knowledge using, respectively, released items from the Massachusetts Test for Educator Licensure (MTEL) and items from the Mathematical Knowledge for Teaching instrument (MKT; H. C. Hill, Rowan, & Ball, 2005).

The spring questionnaire captured teachers' knowledge of students as described below. The teacher questionnaire response rate exceeded 95% at all six data collection points.

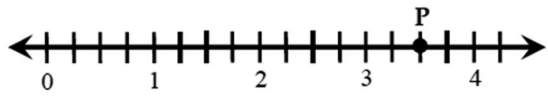
The project also digitally recorded up to three mathematics lessons per school year, for up to nine lessons per teacher. Video capture involved the strategic placement of three unstaffed camera units in teachers' classrooms to maximize the field of vision for recordings; teachers wore a microphone for audio capture during recordings, with a second microphone centrally placed in the room to capture student talk. These dual microphones captured nearly all whole-class discussions clearly, and the teacher microphone consistently captured teacher-student small-group consultations. Teachers chose when to record lessons, though project staff asked teachers not to record lessons on days of testing or test preparation. Lessons averaged approximately 1 hour in length. In total, the project captured and scored 1,713 lessons; due to scheduling constraints (e.g., teachers exiting for medical leave) and technology issues, teachers averaged slightly fewer than three recorded lessons for each year of participation.

Participating teachers' students responded to questionnaires in the spring semester of each school year. These questionnaires contained 26 questions from the TRIPOD survey, an instrument designed to elicit students' perceptions of their mathematics classrooms (Ferguson, 2012). Across all 3 academic years, the study collected student questionnaires from 94% of the students verified as belonging to a project classroom in the spring semester.

Participating teachers' students also completed a project-developed mathematics test in the spring semester of each school year. Project staff designed this assessment to include more cognitively challenging and mathematically complex problems than those found on many state standardized tests (see below for sample items). In doing so, the staff hoped that the assessment would prove more reflective of current standards for student learning (i.e., Common Core Standards for Mathematics) and would more strongly align to the study's mathematics-specific observational and knowledge measures. Nevertheless, this was a broad measure intended for use across multiple projects, rather than solely to test teacher's knowledge of common student misconceptions. Across the 3 academic years, 95% of the students verified as belonging to a project classroom in the spring semester completed the project assessment. Teachers did not receive student performance data. Student scores on the test were estimated using a two-parameter logistic item response theory (IRT) model (Hickman, Fu, & Hill, 2012). Cronbach's alpha for this assessment ranged, across six forms, between .82 and .89.

Finally, each district provided, for all fourth- and fifth-grade students, the following administrative data: mathematics teacher assignment for the duration of the study and up to 2 years prior; student demographic data,

14. Look at the following number line.



What decimal number is represented by point P ?

Figure 2. **Open-ended item on the student project-developed mathematics assessment with a dominant incorrect response (3.2). Source: Hiebert and Wearne, 1983.**

including gender, race, or ethnicity, eligibility for subsidized lunch, English language learner (ELL) status, and special education status; and student scores on state standardized mathematics and English language arts exams.

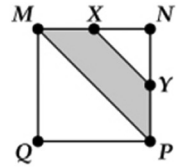
Measures

Project-Developed Mathematics Assessment

During the construction of the project-developed mathematics assessment, we strove to ensure that item distractors (i.e., incorrect responses) matched incorrect thinking patterns commonly held by students. To do so, some of the items drew directly from research on student misconceptions. For instance, one item asked, "Which of the following numbers is greater than 0.23 but less than 0.57?" then posed 0.046 as one of the answer choices. Previous research (Resnick et al., 1989) has established that students know from their study of whole numbers that including a zero to the left of a number does not change the value; overgeneralization of this rule to decimal numbers thus causes errors. This was in fact the case on our assessment as well, where 71% of fourth graders and 61% of fifth graders chose this incorrect option. Open-ended items on the project-developed assessment also strove to elicit common student errors. For instance, based on a misconception noted in Hiebert and Wearne (1983; p. 43), one item asked students to determine the location of point P, marked on a number line at 3 and 2/4s (see Figure 2). The dominant incorrect answer was 3.2, indicating that students had counted over two tick marks rather than thinking of the tick marks as each representing one-fourth.

Other student test items drew from what can roughly be called "craft knowledge," or test-writers' knowledge of the mistakes children make repeatedly with particular content. For instance, one item displayed a square

20. A portion of square *MNPQ* is shaded. Point *X* is the midpoint of side *MN*, and point *Y* is the midpoint of side *NP*.



What fraction of the square shown above is shaded?

A. $\frac{1}{2}$

B. $\frac{1}{4}$

C. $\frac{3}{8}$

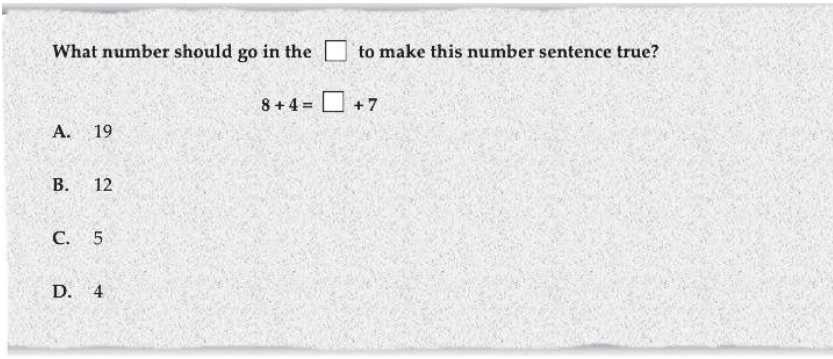
D. $\frac{3}{4}$

Figure 3. **Multiple-choice item on the student project-developed mathematics assessment with a dominant incorrect response (B).**

divided into three nonequal parts (see Figure 3). We constructed distractors based on our sense that students would draw a line parallel line XY in the lower half of triangle MQP, not notice that the parts of the square were not equal, and answer ¼, which was in fact the dominant incorrect response. Though these are not misconceptions found in the research literature, they were familiar to test writers who had long experience working with children.

Teachers' Knowledge of Students

As noted above, the spring teacher questionnaires contained questions intended to assess two different aspects of teachers' knowledge of students. One, knowledge of student misconceptions (KOSM), reflects a component contained within Shulman's (1986, 1987) pedagogical content knowledge and Ball and colleagues' (2008) KCS categories. To measure KOSM, we followed the strategy used by Sadler and colleagues (2013). Specifically, the questionnaire reprinted items from the project-developed mathematics test and asked teachers, "Which [of the following options for this item] will be the most common incorrect answer among fourth [or fifth] graders **in general**?"[2] Using pilot data, project staff selected the student test items to place on the questionnaire strategically, prioritizing items for which the dominant



What number should go in the ☐ to make this number sentence true?

$8 + 4 = \square + 7$

A. 19

B. 12

C. 5

D. 4

The correct answer to this problem is C. a. Approximately what percentage of **your students** being tested today will choose the correct answer?	%
b. Approximately what percentage of fourth grade students in **your district** will choose the correct answer?	%
c. Which will be the most common incorrect answer among fourth graders **in general**? (Please circle **ONE** answer.)	A B D

Figure 4. **Example item on spring teacher questionnaire used to assess both accuracy and KOSM.**
Source: Knuth, Stephens, McNeil & Alibali, 2006.

student response was well established in the research literature and attempting to exclude items that did not meet Sadler et al.'s (2013) definition of a "strong misconception" — i.e., that the most common incorrect student response garnered more than 50% of all incorrect student responses.

Figure 4 shows the response format for these items and provides another sample item that draws from a student misconception well documented in the research literature (Knuth, Stephens, McNeil, & Alibali, 2006). Total, the spring questionnaires included 21 fourth-grade and 20 fifth-grade KOSM items distributed roughly equally across 2 study years (2010–2011; 2011–2012). Five of the 41 KOSM items did not meet Sadler et al.'s (2013) definition of a strong misconception. When excluding these items from the construction of the KOSM measure, results did not change.

To generate KOSM scores for our analyses, we first compared teachers' responses to each question to the actual modal incorrect response of fourth or fifth graders, generating a match/not match indicator.[3] We then estimated the following one-parameter logistic IRT model within grade (as knowledge

of student items differed across grades) using the gsem command in Stata (version 13.1):

$$P(y_{it}=1|\theta_t\alpha_i)=logistic(m\theta_t-\alpha_i) \quad (1)$$

In Equation 1, y_{it} indicates whether teacher t correctly predicted the modal incorrect response among students for item i of the project-developed mathematics test, controlling for α_i, the item difficulty. From Equation 1, we recovered each teacher's "career" (multiyear) KOSM score, θ_t, generated using responses to all KOSM items. We also estimated Equation 1 within school year and grade to recover within-year KOSM scores to use in several analyses detailed below.

We modeled the second measure of teachers' knowledge of students on educational psychologists' notions of judgment *accuracy*, or the extent to which teachers can predict student performance on specified material. To measure this construct, we used the same student items as for KOSM as well as new items in the third study year (2012–2013), this time asking teachers, "Approximately what percentage of **your students** being tested today will choose the correct answer [for this item]?"[4] Both fourth- and fifth-grade teachers answered 37 such items total, with items distributed roughly equally across the 3 years of the study. We argue that this measurement strategy is cognitively simpler and less time-consuming than prior attempts, which asked for prospective scores for each student or each student-item combination. Thus teachers' knowledge of students, as measured here, refers to teachers' knowledge of class performance as a whole, rather than of individual students.

To generate accuracy scores for our analyses, we calculated the actual percentage of correct student answers for each item, addressed potential ceiling and floor effects by transforming both the teacher-predicted and actual percentages into logits, and then differenced the two values.[5] We then used the absolute values of these differences in the following multilevel equation:

$$y_{it}=\beta_0+\alpha_i+\theta_t+\varepsilon_{it} \quad (2)$$

The outcome in Equation 2 represents this absolute difference between predicted and actual logits on item i for teacher t. The model also includes a vector of item fixed effects, α_i, capturing differences in item difficulty to correct for the mix of items taken by a specific teacher, and teacher random effects, θ_t representing teachers' underlying accuracy scores.[6] In addition to estimating accuracy scores within grade from items across all years of the study ("career" scores), we also estimated Equation 2 within school year to recover within-year scores for analyses. We multiplied all scores for the accuracy measure by −1 so that higher scores reflected more accurate predictions.

The use of accuracy and KOSM scores in analyses required that we take additional issues into account, as prior work suggested that classroom features influence teachers' knowledge of students. For example, Klieme and colleagues (2010) showed that teachers are more accurate when predicting the performance of high-achieving, highly motivated students. Hochweber, Hosenfeld, and Klieme (2014) showed that grades given by teachers in more disciplined classrooms better approximate students' test scores than grades given by teachers in unruly classrooms. Relatedly, Martínez and colleagues (2009) found that teachers' assessments of ELL students related more poorly to student test scores than teachers' assessments for non-ELL students. Further, if teachers are generally overoptimistic with regard to student performance on test items (see, e.g., Carpenter et al., 1988, p. 396), teachers with more poorly performing students will be by default less accurate in their predictions. We argue that this artifact of measurement should be controlled, both when constructing accuracy scores and also when modeling student outcomes; doing so helps to remove construct-irrelevant variance and to rule out the possibility that accuracy serves as a proxy for unmeasured student characteristics.

In fact, our data did show teachers to generally overestimate student performance on test items and that teachers' knowledge of student scores correlated with some classroom features (see Table 1). Teachers who instructed higher proportions of Black and ELL students scored lower on both measures. Conversely, those who taught classrooms with higher-achieving students performed better on the accuracy measure. Findings for a TRIPOD-based student report of classroom management (student-level $\alpha=0.61$) ran contrary to expectations, as the composite did not predict either metric. Based on these results, we used adjusted teacher scores in our models, recovered after estimation of the following model:

$$\theta_{gt}=\beta_0+\omega C_{tg}+\varepsilon_{tg} \tag{3}$$

The outcome of Equation 3 represents either teacher t's accuracy or KOSM score while teaching grade g. C_{tg} represents controls for the following classroom-level variables: student race, measured as the proportion of all students taught by teacher t while in grade g during his or her participation in the study who are Black, Asian, Hispanic, or identify with other races or ethnicities; the proportion of teacher t's students in grade g who are ELL students; the average prior state and project-developed mathematics test performance of these students; and teacher t's score on the student survey's classroom management scale.

Other Measures of Teacher Knowledge

Because prior research has demonstrated that teacher content and specialized content knowledge relates to student outcomes (H. C. Hill et al.,

Table 1
Correlations Between Knowledge of Students and Classroom Features

	Accuracy	KOSM
Proportion of students who are Black	−.18**	−.07
Proportion of students who are Asian	.09	.00
Proportion of students who are Hispanic	−.09	−.07
Proportion of students who are of other race/ethnicities	.13*	.06
Proportion of students who are English language learners	−.17**	−.15*
Average prior state mathematics test performance	.27***	.08
Average prior project mathematics test performance	.25***	.10~
Average student-perceived classroom management score	.05	−.09

Note. Number of teachers is 284. Number of teacher-grade combinations is 292.
~$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$.

2005; Rockoff, Jacob, Kane, & Staiger, 2008) and because such knowledge may itself facilitate the noticing of student performance and student misconceptions and thus independently influence instruction (see Figure 1), we used our MKT and MTEL items to create a measure of mathematical knowledge. A confirmatory factor analysis of teacher performance on these items (Charalambous, Hill, McGinn, & Chin, 2017) showed a one-factor model adequately fit the data, and thus we constructed a single metric encompassing both item types (*MKT/MTEL*); teachers' scores on this metric had an estimated marginal reliability of .92.

Measures of Instructional Quality

Lessons were scored on the Mathematical Quality of Instruction (MQI) (H. C. Hill, Blunk, et al., 2008), which contains items related to teachers' work with student thinking. By comparing Figure 1's instructional processes to the available metrics on the MQI, we hypothesized that teachers' accuracy and KOSM would be related to teacher remediation of student mistakes (*remediation*) and the degree to which teachers incorporate student thinking into lessons (*use of student productions*). We hypothesized that teachers with more knowledge of student thinking (both KOSM and accuracy) would score more highly on both MQI items. Clearly, there are many more ways that teachers' knowledge of students may influence instruction, including (as noted in our introduction) choosing or designing tasks, forming appropriate instructional groups, and adjusting lesson pacing. These constructs are not readily observable from video, however, and thus not measured in our study. This makes our measures of instruction an only partial test of the hypothesis that stronger knowledge of students correlates with higher-quality instruction.

To obtain teacher-level scores for these items, each recorded lesson was scored in 7.5-minute segments by two raters who were screened, certified, and trained by the study. We settled on 7.5 minute segments, as opposed to longer or shorter segments, because raters reported that longer segments were too cognitively burdensome and because shorter segments meant significantly more scoring time and cost. Every 7.5-minute segment in each lesson was scored, as were final segments that were more than a minute long. Exact rater agreement was 66% and 76%, respectively, for remediation and use of student productions.

A factor analysis revealed that remediation and use of student productions failed to form a single, coherent factor. Instead, we leveraged the fact that raters scored each item on average 101 times per teacher, across diverse lessons, days, content, and pedagogies, to arrive at an estimate of their underlying propensity to engage in each activity. Notably, prior generalizability studies (e.g., H. C. Hill, Charalambous, & Kraft, 2012) have demonstrated that our scoring structure (i.e., multiple raters scoring all segments of multiple lessons for each teacher) can result in scores that reasonably distinguish teachers from one another on similar MQI items, even after taking into consideration these other influences.

To generate scores on each item for each teacher, we first averaged scores across segments and raters to the lesson level. We then estimated the following multilevel model, in which lessons are nested within teachers:

$$y_{lt}=\beta_0+\theta_t+\varepsilon_{lt} \tag{5}$$

The outcome, y_{lt}, represents teacher t's score for either remediation or use of student productions on lesson l. The model also contains teacher random effects, θ_t, which reflects teacher t's underlying MQI score. Using empirical Bayes (Raudenbush & Bryk, 2002), teacher scores were adjusted to account for differences in reliability caused by varying numbers of lesson-level MQI scores included in the model. Intraclass correlations (ICCs) were .53 for teacher remediation scores and .66 for use of student production scores. These ICC estimates adjust for the average number of lessons observed across teachers so that they reflect the amount of variance attributable to differences between typical teachers on all their scores in our sample, as opposed to the variance attributable to differences between teachers for a single observation score. The typical teacher was observed on just under six lessons.

Finally, we used teachers' KOSM and accuracy scores to predict *monitoring, evaluation, and feedback*, a measure created from students' answers to four TRIPOD items regarding topics such as whether their teacher knows when the class does or does not understand material, and whether their teacher checks to make sure that the class understands what is being taught (α=0.70). To estimate a score for each teacher, we first averaged responses

across these four items for each student, then estimated the following multilevel model separately within grade (to account for the fact that teachers' classroom quality, as determined by students' perceptions, may vary by the grade taught), where students are nested within years (i.e., a teacher-year interaction, or "classroom," effect), which are nested within teachers:

$$y_{syt} = \beta_0 + \theta_t + \nu_{yt} + \varepsilon_{syt} \quad (6)$$

The outcome, y_{syt}, represents the average of the responses to the four monitoring, evaluation, and feedback items from student s in year y taught by teacher t. The model also contains classroom random effects, ν_{yt}, and teacher random effects, θ_t; the latter captures teacher t's underlying score on the construct. Using empirical Bayes (Raudenbush & Bryk, 2002), teacher scores were adjusted to account for differences in reliability caused by variability in class size.

Analysis Strategy

This paper seeks to understand to what extent teachers' accuracy and KOSM scores are stable across cohorts of students, and how well scores differentiate among teachers; the extent to which these measures relate to each other and other similar measures of teacher knowledge; whether teachers' accuracy and KOSM scores predict instruction; and how well teachers' scores predict student outcomes. We outline a strategy for answering each question below.

Score Stability and Differentiation Among Teachers

We estimated three reliability metrics, one of which focused on measure stability. For the KOSM measure, we used estimates of the marginal reliability produced in the IRT model described above. The marginal reliability statistic compares the variance of teacher scores to the expected value of error variance of scores and is comparable to ICCs of classical test theory (see Sireci, Thissen, & Wainer, 1991). For the accuracy measure, we estimated the signal-to-noise ratio in teacher scores using the ICC statistic, adjusted for the average number of items answered by teachers within each grade. Finally, we also used the within-year estimates of accuracy and KOSM scores to examine cross-year correlations, a measure of consistency. Results are presented separately by grade, as the student items used were specific to each grade level. For this analysis, we used the largest sample of teachers possible; that is, the sample of all teachers who responded to the knowledge of student items on any spring questionnaire, 315 and 306 teachers for accuracy and KOSM, respectively.

Relationships among related knowledge measures. Ideally, we would have followed Krauss et al. (2008) and conducted a factor analysis of items representing teachers' KOSM, judgment accuracy, and MKT/MTEL. However, because scores were generated from different data collection instruments and calculated using different procedures, we thought it likely that any factors identified would relate to method rather than knowledge type. Instead, we performed simple correlations of scores and used correlations found in prior research to interpret our results.

Predicting instructional quality. We used the KOSM and accuracy scales to predict teachers' instructional quality as captured in the MQI remediation and use of student productions scales and the TRIPOD's monitoring, evaluation, and feedback scale. Specifically, we performed a simple ordinary least squares regression:[7]

$$y_t = \beta_0 + \omega\theta_t + \psi\pi_t + kC_t + \varepsilon_{syt} \quad (7)$$

In this model, we controlled for teachers' mathematical knowledge (π_t) because this knowledge may independently influence instructional quality. We also controlled for student characteristics (C_t), based on both the relation of teacher scores to student characteristics described above, and upon our sense that prior achievement and related characteristics may influence teachers' opportunities to listen to and respond to student thinking. The coefficient ω captures the relationship between teacher KOSM and accuracy and teachers' scores on the instructional outcomes, y_t. One caveat to this analysis involves the directionality of the relationship between teachers' knowledge of students and their instructional practice. Specifically, teachers' remediation and use of student productions may increase their KOSM and judgment accuracy, as they may learn more student thinking from these activities. Unfortunately, a test of the direction of the relationship (i.e., predicting prior and/or future instructional quality with knowledge) would be underpowered in our sample due to missing-by-year data.

Predicting student outcomes. Many researchers and policymakers operate under the assumption that teachers' knowledge of students works to improve student outcomes. We could not test this hypothesis in a causal manner as we did not randomly assign students to teacher knowledge levels, nor did we have instructional measures that specifically match all the theoretical mediators. However, we examined associations between these forms of teacher knowledge and student outcomes in models designed to limit the bias in estimates due to teacher and student sorting to classrooms. Our basic

multilevel equation, where students are nested within years nested within teachers, was:

$$y_{spcgyt}=\beta_0+\alpha X_{sy-1}+\delta D_{sy}+\phi P_{pcgyt}+kC_{cgyt}+\eta+\omega\theta_{gt}+\psi\pi_t+\mu_t+\nu_{yt}+\varepsilon_{spcgyt} \quad (8)$$

where the outcome, y_{spcgyt}, represents the test performance on either the state standardized or project-developed mathematics test of student s, in classroom p, in cohort (i.e., school, year, and grade) c, taking the test for grade g, in year y when project data was collected (i.e., 2010–2011, 2011–2012, or 2012–2013), taught by teacher t. To minimize bias, Equation 7 contains the following:

- X_{sy-1}, a vector of controls for student prior test performance;
- D_{sy}, a vector of controls for student demographic information;
- P_{pcgyt}, classroom-level averages of X_{sy-1} and D_{sy} to capture the effects of a student's peers;
- C_{cgyt}, cohort-level averages of X_{sy-1} and D_{sy} to capture the effect of a student's cohort;
- η, school and grade-by-year fixed effects;
- θ_{gt}, accuracy and KOSM scores for teacher t for grade g, adjusted for classroom features;
- π_t, teachers' MKT/MTEL scores;
- μ_t, a random effect on test performance for being taught by teacher t; and
- ν_{yt}, a random effect on test performance for being taught by teacher t in year y.

Although this model is similar to those frequently used by states, districts, and research studies to obtain value-added scores for teachers, the controls for classroom- and cohort-level averages of prior test performance and demographics and the school fixed effects are unique. We argue that using this heavily controlled model helps address the observational nature of our analyses (see, for example, Rothstein, 2009). The classroom aggregates and school fixed effects also help control for any remaining bias in teachers' adjusted accuracy and KOSM scores after correcting those scores for classroom characteristics as described above. As an additional check on our findings, however, we also predicted student performance on the state test in the year before teachers in our sample participated in the National Center for Teacher Effectiveness project. Others (e.g., Kane & Staiger, 2012; Kane, Taylor, Tyler, & Wooten, 2011) have used such a check to rule out the possibility that student composition biases teachers' classroom-based scores. Finally, to disentangle the influence of teachers' knowledge of students from related predictors and to help avoid omitted variable bias, our multilevel model also contained teachers' MKT/MTEL scores.

After completing our main investigation into the relationship between teachers' career knowledge of student scores and student performance, we also investigated the relationship between within-year teacher scores

and student performance, testing the possibility that these knowledge measures may be composed of both a "stable" trait (i.e., general knowledge of students) and a year-specific deviation (i.e., knowledge of a particular group of students). We note, however, that the reliability of within-year scores was lower as we estimated such scores using fewer items. Furthermore, because we did not measure KOSM in 2012–2013, this within-year analysis includes only students taught by the project teachers in 2010–2011 and 2011–2012, reducing the student sample to 7,785 students.

Next, we examined the association between teachers' career knowledge of students' scores on student test performance for students at varying levels of prior test performance. This analysis followed results found by Sadler and colleagues (2013), who noted heterogeneous effects of teachers' knowledge on outcomes for groups of students stratified based on pretest scores. Finally, we explored whether the valence—positive or negative—of teachers' predictions of students' performance related to students' actual test performance. To do so, our student outcomes models included indicators for teachers who markedly over- or underpredicted their students' performance on project-developed test items (i.e., greater or less than one *SD* from the mean score). Neither indicator captured absolute tendencies but instead reflected a teacher's expectations relative to his or her peers, after controlling for features of the teacher's classroom.

Results

Table 2 provides teacher-level summary statistics for key measures in our analyses. We could not, as per guidelines for using the MKT measures, publish the percent correct for this metric. However, from the table, other facts emerge. First, the average teacher correctly guessed the most common wrong answer 55% of the time. Second, the mean difference between teachers' predictions and students' actual performance across all items was almost 28 percentage points (0.277). Although students gave their teachers high marks on the monitoring, evaluation, and feedback metric (mean = 4.416), video observers recorded few instances of remediation and use of student productions. Specifically, the average teacher scores for both metrics (1.370 and 1.218) corresponded with only 32 and 18 percent of rater-scored segments scoring above 1 ("not present") on these items, respectively. For more information on teacher performance on our KOSM and judgment accuracy measures, please see the online appendix. To simplify interpretation of the results from our analyses, these measures were standardized to have a mean of zero and unit variance.

Score Stability and Differentiation Among Teachers

Using adjusted ICCs, we estimated the reliability of accuracy scores for the typical fourth grade teacher to be .74 and fifth-grade teacher to be .72.

Table 2
Raw Summary Statistics of Measures of Knowledge and Instructional Quality

	Mean	*SD*	Min	Max	Scale
KOSM	0.553	0.146	0.111	0.895	Dichotomous
Accuracy	0.277	0.080	0.109	0.587	[0,100]
Remediation	1.370	0.162	1	2.055	Ordinal [1,3]
Use of student productions	1.218	0.156	1	1.718	Ordinal [1,3]
Monitoring, evaluation, and feedback	4.416	0.204	3.601	4.853	Ordinal [1,7]

Note. Number of teachers is 284.

We also investigated the adjusted ICCs for the set of 22 fourth-grade teachers and 25 fifth-grade teachers who responded to all items ($N=37$) measuring accuracy. The adjusted ICCs for these samples were .76 and .77, respectively. These values suggest that, with enough item responses, our measure of accuracy can differentiate among teachers' performance on this construct.

The marginal reliability statistic produced following the estimation of the KOSM IRT model was .21 for fourth-grade teachers and .40 for fifth-grade teachers. The magnitude of the average standard errors of scores reflected these reliability coefficients, suggesting high imprecision for the average individual; for fourth-grade teachers, the average magnitude of the standard error of KOSM scores was .93 *SD*, and for fifth grade, the average magnitude was .85 *SD*. The low score reliability estimates suggested that the KOSM measure did not adequately differentiate teachers.

Constructing and then correlating within-year scores provided evidence on the stability of this trait in teachers. Across both grade levels and combination of years, we found moderate cross-year correlation of accuracy scores, ranging from .29 to .53; these estimates suggested that teachers' ability to predict the proficiency of their students was somewhat consistent from school year to school year, despite changes in the students taught. Furthermore, correlations were in the same range as the cross-year correlations of other measures of teacher quality (Goldhaber & Hansen, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Polikoff, 2015). KOSM scores, as expected given their estimated overall reliability, demonstrated less consistency from each year to the next. For fourth-grade teachers, scores correlated at .22 between 2010–2011 and 2011–2012; for fifth-grade teachers, this correlation was slightly higher, at .26.

The differences in score reliability across these two measures might have emerged for several reasons. The accuracy questions were asked on three different spring surveys, and the KOSM questions on just two, rendering

roughly one third fewer items for estimating the latter. The cross-year correlations for the KOSM measure may have also been lower because of changes to the language of questions assessing the construct between the 2010–2011 version and the 2011–2012 version of the questionnaire. Finally, even though we used pilot student test data to select items based on wide gaps in distractor endorsements, on the actual assessments, five of the 41 KOSM items failed to meet Sadler et al.'s (2013) bar for having a strong misconception elicited; such items tapping more diffuse sets of misconceptions may fail because such knowledge is neither systematically held by nor useful to teachers, resulting in a less reliable score overall. Whatever the reason, we did not meet with success in our measurement strategy for KOSM. Nevertheless, we proceeded with using teacher KOSM scores in the analyses below because of KOSM's role as a central construct of interest in both our work and prior work. We note, however, that low score reliability tends to bias observed relationships toward zero, and take this into account in our discussion below.

Relationships Among Related Knowledge Measures

Like Carpenter and colleagues (1988), we find a very weak relationship between teachers' knowledge of student thinking (KOSM) and judgment accuracy, $r(284) = .098$, $p < .01$, suggesting that these are relatively disparate knowledge constructs. We find modest relationships between the knowledge of students measures and teachers' mathematical knowledge (KOSM, $r(284) = .13$, $p < .05$; accuracy, $r(284) = .245$, $p < .001$).

Predicting Instructional Quality

Next, we investigated whether our teacher accuracy and KOSM scores predicted instructional quality as measured by both the MQI and TRIPOD.

Table 3 shows some evidence for the importance of judgment accuracy, in particular, in predicting teachers' instructional practice, even when controlling for classroom characteristics and teacher content knowledge. Specifically, we see that accuracy positively predicted teachers' use of student productions in models that both include and exclude MKT/MTEL scores. This finding converges with intuition, as active recognition and incorporation of student mathematical thinking in the classroom should provide teachers with additional information on their students' content mastery, and information on students' content mastery likely assists teachers in asking appropriate questions and using their thinking. Accuracy also predicted teachers' remediation of students' mathematical errors during class when excluding the control for MKT/MTEL. When including this control, however, the relationship remained positive but attenuated; the primacy of

Table 3
Predicting Teachers' Instructional Quality Using Career Knowledge of Student (KOS) Scores

	Video-based instruction measures				Student-based instruction measures	
	Remediation		Use of student productions		Monitoring, evaluation, and feedback	
Accuracy	.137*	.078	.214***	.176**	.086	.072
	(.059)	(.059)	(.057)	(.058)	(.053)	(.055)
KOSM	.093	.064	−.039	−.057	−.031	−.037
	(.059)	(.059)	(.057)	(.057)	(.053)	(.054)
MKT/MTEL		.254***		.165**		.058
		(.061)		(.060)		(.056)
N	284	284	284	284	292	292

Note. All models include controls for demographic and academic characteristics of teachers' students. Models with the monitoring, evaluation, and feedback instructional quality measure as the outcome are at the teacher-grade level. ~$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

MKT/MTEL's impact in this model is perhaps unsurprising, given the importance of content knowledge for identifying errors in the first place.

No significant relationships emerged between teachers' ability to identify student misconceptions and either video-based instruction measure, though the direction and approximate magnitude of the relationship between KOSM and remediation matched our intuition that teachers with stronger KOSM would also engage in more in-class remediation during observed lessons.

Finally, neither knowledge of students metric was found to be related to students' TRIPOD reports of teacher monitoring, evaluation, and feedback; this may reflect the low internal consistency of the latter (α = .59).

Predicting Student Outcomes

As noted above, both theory and prior empirical evidence suggest that accuracy and KOSM scores should relate to student test performance. The results of our first analyses, using career scores derived from all years of survey data, appears in the first two columns of Table 4.

From this table, we see teachers' career KOSM scores showed a small negative relationship to their students' performance on the project-developed test and a larger positive relationship to their students' performance on the state test; neither point estimate, however, was significant.

Table 4
Predicting Student Math Test Achievement Using KOS Scores

	Career KOS scores		Within-year KOS scores	
	Project	State	Project	State
Accuracy	.023*	.021~	.035**	.025~
	(.011)	(.012)	(.011)	(.012)
KOSM	−.009	.016	−.003	.023~
	(.010)	(.011)	(.011)	(.012)
Number of students	9,636	9,636	7,785	7,785

Note. Number of teachers is 284. All models include controls for student prior test performance, grade-year interaction fixed effects, student demographics, classroom-level aggregates, cohort-level aggregates, school fixed effects, teacher random effects, and teacher-year interaction random effects, as well a control for teacher career MKT/MTEL scores. ~$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

The situation is different for accuracy. The table shows that teachers' scores in this realm positively related to their students' performance on the project-developed mathematics test, even when controlling for factors that might bias this relationship, such as classroom- and cohort-level aggregates of student prior test performance, student demographic characteristics, and school fixed effects. The magnitude of the coefficient for teachers' accuracy scores was .023; because all knowledge of student measures are standardized to have a mean of zero and unit variance, we can interpret this coefficient as the difference in student test scores (which are themselves student-level *Z*-scores) associated with a one-*SD* difference in teacher accuracy. To facilitate interpretation of the coefficient's magnitude, consider that fourth graders improved on average 0.59 *SD* on the project-developed test over the course of an average school year, and that fifth graders improved on average .48 *SD*; thus, students taught by a teacher one *SD* above average in terms of accuracy gained approximately 1.5 weeks of schooling compared with those taught by an average teacher. The size of the accuracy estimates is also slightly smaller than but still in league with that of other variables in economics of education research, such as teaching experience (.08 *SD* for experience above 5 years; Goldhaber & Hansen, 2010; Kane, Rockoff, & Staiger, 2008; Papay & Kraft, 2015), the effect on students' math scores of being taught by Teach For America teachers (0.024 to 0.10 *SD* in Kane et al., 2008; Clark et al., 2013; Xu, Hannaway, & Taylor, 2011), and the average impact of educational interventions (0.07 *SD*; C. J. Hill, Bloom, Black, & Lipsey, 2008).

We also investigated the association between teachers' accuracy and state standardized test performance. Doing so helps alleviate a concern

with the above analysis: that the same items were used to measure both teachers and students' performance, albeit in different ways. As the second column of Table 4 shows, we saw slightly weaker but still positive and marginally significant relationship between accuracy scores and student state test performance ($\beta = .021$, $p \leq .10$). The consistent positive relationship between teachers' accuracy scores and their students' performance on different tests corroborates hypotheses posited by prior literature.

The final two columns of Table 4 further support our overall conclusions regarding the importance of accuracy for student outcomes and provide suggestive evidence for the importance of KOSM as well. In these columns, we used within-year scores to explore the possibility that the knowledge of student measures may capture an important year-specific component of teacher capability. Again, we found that teacher accuracy scores significantly and positively predicted student test performance on both the project-developed test and the state standardized test, with stronger associations demonstrated on the project-developed test. For KOSM, within-year scores also predicted student performance on the state mathematics test, despite our failure to find a significant relationship between the career estimate of KOSM to either outcome. This difference appeared to be driven by differences in the student sample used in each multilevel regression; for example, when we used career knowledge of student scores to predict the state test performance of the same 7,785 students used in the within-year models, the regression coefficient for career KOSM increased to a similar magnitude. Our within-year analysis thus suggested two conclusions: Teachers' knowledge of a particular group of students predicts test performance, as does teachers' knowledge of students more generally, and KOSM, though contributing less than accuracy to outcomes, may still be an important capability. However, this effect surprisingly appeared only for the state test; its relationship to the project-administered test, from which the teacher KOSM items were based, was zero.

Table 5 depicts results from our effort to ascertain whether measurement issues, and specifically the artifact of measurement that arises when teachers of lower-performing students generally overestimate student test performance, drove the results on accuracy. To address this issue, we used teacher scores to predict student performance on the state test in the year *before* participation in the project; doing so is thought to rule out bias brought about by classroom compositional factors (Kane & Staiger, 2012; Kane et al., 2011). Our analysis showed that teacher accuracy scores still significantly predicted state mathematics test performance of the students taught by the subset of 185 teachers from our main analyses that appeared in academic years both before and during the time of the project.[8] Specifically, we show in Table 5 that associations between teachers' accuracy scores and mathematics achievement were of similar magnitude in both time periods.

Table 5
Predicting Student State Math Test Achievement Using Career KOS Scores

	Project participation years	First year prior to project participation
Accuracy	.037*	.043**
	(.016)	(.016)
KOSM	.030*	.012
	(.015)	(.017)
Number of students	6,565	3,906

Note: Number of teachers is 185. Models include controls for student prior test performance, grade-year interaction fixed effects, student demographics, classroom-level aggregates, cohort-level aggregates, school fixed effects, teacher random effects, teacher-year interaction random effects (except for the "first year prior" model), as well as a control for teacher career MKT/MTEL scores. ~ $p < .10$,
*$p < .05$, **$p < .01$, ***$p < .001$.

Table 6 shows the results from our exploration into a finding reported by Sadler and colleagues (2013). Similar to Sadler et al., we found that the association between teachers' knowledge of students and student outcomes differed across student populations. Specifically, we found suggestive evidence of larger associations between accuracy and student outcomes for students who perform better at baseline. Coarsening the data by categorizing students into terciles of baseline performance suggested this interaction to be mainly driven by a lack of effect of accuracy for students in the lowest tercile. Though one possible explanation for this particular result is that lower-performing students are less able to profit from instructional resources such as teacher knowledge, similar to the observation that some students are unresponsive to even intensive educational interventions (e.g., Toste et al., 2014), our finding for KOSM contradicted this possibility. For the latter metric, the effect of teacher knowledge on outcomes was greater for students with lower baseline achievement. One possibility for the observed heterogeneous effects of teachers' knowledge of students may be that teachers differentially engage in the instructional practices that mediate the relationship between knowledge of students and outcomes. As noted earlier, however, we lack data on all such practices, preventing us from providing satisfactory empirical evidence that would support hypothesis testing.

Finally, Table 7 shows the association between the valence of teacher predictions and student test performance. We undertook this analysis to explore whether teachers' expectations regarding student performance relate to those students' eventual performance. We found no evidence that students taught by teachers who had a tendency to underpredict or overpredict performance performed differently on both the project and state mathematics tests, after controlling for their teachers' accuracy score. We also note that the point estimate for teachers' accuracy on student test performance for

Table 6
Heterogeneously Predicting Student Test Performance with Career KOS Scores

	Project		State	
	Continuous	Categorical	Continuous	Categorical
Accuracy	.024*	.028~	.023~	.028*
	(.011)	(.015)	(.012)	(.015)
Accuracy * prior test performance	.013~		.027***	
	(.008)		(.007)	
Accuracy * 1st tercile prior test performance		−.024		−.032*
		(.016)		(.013)
Accuracy * 2nd tercile prior test performance		omitted		omitted
		(.)		(.)
Accuracy * 3rd tercile prior test performance		.012		.019
		(.017)		(.014)
KOSM	−.009	−.012	.015	.009
	(.010)	(.013)	(.011)	(.014)
KOSM * prior test performance	−.015*		−.006	
	(.007)		(.007)	
KOSM * 1st tercile prior test performance		.024		.023~
		(.015)		(.013)
KOSM * 2nd tercile prior test performance		omitted		omitted
		(.)		(.)
KOSM * 3rd tercile prior test performance		−.014		−.005
		(.015)		(.013)

Note: Number of teachers is 284. Number of students is 9,636. All models include controls for student prior test performance, grade-year interaction fixed effects, student demographics, classroom-level aggregates, cohort-level aggregates, school fixed effects, teacher random effects, teacher-year interaction random effects, as well as a control for teacher career MKT/MTEL scores. ~ $p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

the state test remained significant even after controlling for teachers' prediction valence; the point estimate for this measure on project-test performance slightly attenuated, but the standard error also increased due to the inherent covariance between teacher valence and accuracy controls in the model. The robustness of accuracy supports the notion that the effect of teacher accuracy on outcomes is not confounded with the effect of teachers' expectations, which themselves may instigate changes in instructional behaviors that lead to differential outcomes for students.

Conclusions

This article provides evidence regarding the importance of teachers' knowledge of students. For judgment accuracy, our investigation produced

Table 7
Predicting Student Math Test Achievement using Career KOS Scores and Valence

	Project	State
Accuracy	.018	.040*
	(.016)	(.017)
Overpredictor	−.019	.067
	(.043)	(.046)
Neither	omitted	omitted
	(.)	(.)
Underpredictor	−.011	−.042
	(.035)	(.039)
KOSM	−.006	.018
	(.010)	(.011)

Note: Number of teachers is 284. Number of students is 9,636. All models include controls for student prior test performance, grade-year interaction fixed effects, student demographics, classroom-level aggregates, cohort-level aggregates, school fixed effects, teacher random effects, teacher-year interaction random effects, as well as a control for teacher career MKT/MTEL scores. ~ $p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

evidence that teachers' accuracy is an identifiable teacher-level trait; teacher scores remained stable over time, and showed acceptable levels of reliability, suggesting that teachers were differentiable on the metric. Performance on this metric was also related to teachers' mathematical knowledge and their engagement with specific classroom activities we hypothesized would be related to accuracy, specifically remediation of student misconceptions and use of student productions. In line with prior theoretical expectations, this metric predicted student outcomes, demonstrating significant relationships to performance on the project-administered test from which it was derived and marginally significant relationships in models from our main analyses predicting student performance on state standardized tests. While our analyses are not causal, we controlled for student sorting by using a fairly rigorous model, one incorporating peer- and cohort-level averages of student demographics and baseline test performance in addition to school fixed effects. These results suggest that a construct roughly titled "knowing where your class is, in terms of mastery of content" belongs in contemporary delineations of teacher knowledge.

We see two directions for future research given this finding. First, as discussed below, this correlational evidence suggests that subsequent causal work might be fruitful. Second, mechanisms behind these relationships bear further investigation. As noted above, the directionality between our variables is not completely clear; active recognition and incorporation of student mathematical thinking in the classroom might provide teachers with

additional information on their students' content mastery, thus improving accuracy, and information on students' content mastery likely assists teachers in remediating misconceptions, asking appropriate questions, and using their thinking. More accurate knowledge of what students know and do not know also may assist teachers in other ways, for instance in planning to reteach content that has not been mastered and in designing tasks and instruction that intentionally elicits typical student mistakes with content. Although we measured accuracy at the classroom level, there may be an individual-level component as well, with highly knowledgeable teachers aware of which students are missing which content, which then enables regrouping students for efficient remediation. As our measures in this domain are quite blunt, these are all issues for future research.

The story is more complicated for KOSM, which proved more difficult to measure. Although there was some evidence of cross-year correlation in teacher scores, reliabilities were subpar and no significant relationships to instruction appeared. However, within-year estimates of this metric did predict student state test performance; neither career nor within-year estimates, however, predicted performance on the test from which teacher KOSM scores were derived. We find the inconsistency in predictive power to be somewhat surprising, given the central place of KOSM in most theories of teacher knowledge (Ball et al., 2008; Shulman, 1986), the presumed stability of KOSM as a type of teacher knowledge, and the similarity of our efforts to others in this area (Sadler et al., 2013) although we also note that the two empirical studies of this area returned mixed results (Carpenter et al., 1988; Sadler et al., 2013). One reason might be the difficulty of measuring this domain; we had trouble finding research-based student misconceptions at our grade levels and thus constructed some items from "craft knowledge," perhaps leading to the low observed score reliability and weak connections to practice. It may also be that teachers' knowledge of student misconceptions may not be orderly, in the sense of belonging to a spectrum from poor to strong or novice to expert; instead, teachers' knowledge may accumulate in "bits and pieces" (Shavelson & Kurpius, 2012) from personal experience. If so, this line of reasoning suggests that teachers' knowledge may be strongly context-dependent and thus not make itself amenable to measurement. This suggests that analysts must continue to explore this construct de novo, perhaps using qualitative probes and interviews, and assess whether this construct is worth continuing to invest in, in terms of measurement.

Notwithstanding the mixed results from KOSM, the evidence reviewed here suggests that teacher knowledge of students is an identifiable trait, associated with teachers' content knowledge, instruction, and student outcomes in ways we would expect, given theory (e.g., Shulman, 1986). Notably, we also offer the first evidence of a main (rather than interactional) effect of teacher knowledge of students on student outcomes independent of teacher knowledge of subject matter. Though the association is small, in other

words, teachers' knowledge supports improved outcomes for all students. This strengthens the argument that teachers' knowledge—of content and of students as they learn that content—is an important resource for the production of quality instruction and student learning.

By extension, this suggests that interventions designed to improve teachers' knowledge of student learning may bear fruit; such interventions could also provide causal evidence that these constructs play an important role in teaching. In fact, such past efforts in the area of KOSM (e.g., Carpenter et al., 1989; Jacobs et al., 2007) have proven effective. However, programs in which teachers study student data, and which theoretically should improve teacher accuracy, have returned largely null results on student achievement (Cordray, Pion, Brandt, Molefe, & Toby, 2012; Henderson, Petrosino, Guckenburg, & Hamilton, 2008; Konstantopoulos, Miller, & van der Ploeg, 2013; Quint, Sepanik, & Smith, 2008; West, Morton, & Herlihy, 2016; for a partial exception, see Carlson, Borman, & Robinson, 2011). This presents a puzzle: if teacher knowledge of student performance is related to student outcomes, why do so many of these programs fail?

One possibility is that these data-focused programs may only develop a very narrow knowledge of particular students' performance on particular test items. Instead, accuracy as we measured it may tap a broader form of knowledge, encompassing aspects of noticing and judging student understanding during instruction itself. Another reason may be in the activities teachers engage after receipt of the information about students. In our study, teachers with more knowledge of students engaged in more remediation of student misconceptions and more often used student thinking in instruction. Similarly, empirical evidence regarding professional noticing suggests that teachers who perceive and understand student thinking plan instruction differently (Barnhart & van Es, 2015; Jacobs et al., 2011). By contrast, case studies of teachers engaged in the study of data indicate that they were likely to regroup students and reteach content, but much less likely to change instruction or to diagnose student misconceptions (e.g., Goertz, Olah, & Riggan, 2009). This suggests that interventions designed to improve teachers' knowledge of student performance might focus improving teachers' capacity to notice student performance in situ and to make corresponding adjustments to instruction.

Finally, our findings suggest three paths for future research. First, we know little of how teachers' knowledge of students operates in subject matter beyond mathematics and science, where we and most other scholars have concentrated our efforts. Teachers' knowledge of students may operate differently in other domains. Second, we have only incomplete in-practice measures of what teachers actually *do* with such knowledge. Although we opened this article with evidence from observational and case studies of teaching, and although the noticing literature makes a strong case based on interview and written assessment data, we argue that more can and

should be learned about how knowledge in this arena supports teacher action. Uncovering the specifics of how teachers use such knowledge may take a program of research similar to that built around teachers' content knowledge, where the development of measures in this domain allowed investigators to understand how teachers deployed such knowledge in the classroom. Third, scholars may want to investigate what experiences lead teachers to better knowledge of their students. Although there is an abundance of evidence regarding demographic and classroom-based predictors of teachers' judgment accuracy, for instance, little of this research examines whether factors such as teaching experience, teacher training, pedagogical methods, and assessment practices lead to better teacher accuracy (for an exception, see Martínez et al., 2009). Scholars may also want to propose and test professional development or curricular experiences designed to sharpen teachers' accuracy, similar to the ways in which teachers' knowledge of student misconceptions (and learning more generally) has been the topic of professional development in mathematics (e.g., Carpenter et al., 1989; Philipp et al., 2007).

Notes

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education (Grant R305C090023) to the President and Fellows of Harvard College to support the National Center for Teacher Effectiveness. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

[1]Though these are similar to arguments made during measure validation processes, we have not cast our investigation in a validation framework. One major reason is that, as Kane (2010) emphasizes, one validates *scores*, and validates them for a *particular use*. We do not propose using these scores for making judgments about teachers. Instead, we care whether the construct itself, as measured by our instrument, is of importance to teaching and learning, similar to the approach Baumert et al. (2010) take with teachers' pedagogical content knowledge.

[2]This is the exact wording as it appears on the 2011–2012 spring teacher questionnaire of the study. In 2010–2011, the wording for this question was, "Which [of the following options for this item] do you think will be the most common **incorrect** response among your students?" The study changed to ask about fourth and fifth graders in general in light of low reliabilities observed in the first year of data collection.

[3]Because of the wording differences on the first (2010–2011) survey noted above, we compared teachers' responses to this question to the actual modal incorrect response of his or her students if a modal response existed.

[4]This is the exact wording as it appears on the 2011–2012 and 2012–2013 spring teacher questionnaires of the study. In 2010–2011, the wording for this question was, "What percentage of your students being tested today do you think will choose the correct answer [for this item]?" The wording between years is substantively similar.

[5]In order to estimate the logits of teacher predictions or calculated student percentages of 0% or 100%, we rescaled these values to 1% and 99%, respectively.

[6]Higher difficulties indicated items for which teachers' predictions were further off.

[7]Because the monitoring, evaluation, and feedback instruction measure was estimated at the teacher-grade level, we performed this specific ordinary least squares regression at the teacher-grade level.

[8]Two sample *t*-tests with equal variance showed that teachers included in these analyses did not differ significantly from those excluded in terms of accuracy and KOSM scores.

ORCID iD

Mark Chin https://orcid.org/0000-0002-0136-2068

References

An, S., Kulm, G., & Wu, Z. (2004). The pedagogical content knowledge of middle school, mathematics teachers in China and the U.S. *Journal of Mathematics Teacher Education*, *7*(2), 145–172.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389–407.

Barnhart, T., & van Es, E. (2015). Studying teacher noticing: Examining the relationship among pre-service science teachers' ability to attend, analyze and respond to student thinking. *Teaching and Teacher Education*, *45*, 83–93.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . .Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180.

Bell, C. A., Wilson, S. M., Higgins, T., & McCoach, D. B. (2010). Measuring the effects of professional development on teacher knowledge: The case of developing mathematical ideas. *Journal for Research in Mathematics Education*, *41*(5), 479–512.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74.

Borasi, R. (1994). Capitalizing on errors as "springboards for inquiry": A teaching experiment. *Journal for Research in Mathematics Education*, 166–208.

Carpenter, T. P., Fennema, E., Peterson, P. L., & Carey, D. A. (1988). Teachers' pedagogical content knowledge of students' problem solving in elementary arithmetic. *Journal for Research in Mathematics Education*, *19*(5), 385–401.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C. P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal*, *26*(4), 499–531.

Carpenter, T. P., Moser, J. M., & Romberg, T. A. (Eds.). (1982). *Addition and subtraction: A cognitive perspective*. Mahweh, NJ: Erlbaum.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, *33*(3), 378–398.

Charalambous, C. Y., Hill, H. C., McGinn, D., & Chin, M. (2017). *Content knowledge and teacher knowledge for teaching: Exploring their distinguishability and contribution to student learning*. Manuscript in progress.

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). *The effectiveness of secondary math teachers from Teach For America and the Teaching Fellows Programs* (NCEE Publication No. 2013–4015). Washington, DC: National Center for Education Evaluation and Regional Assistance.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.), *Third handbook of research on teaching* (pp. 255–296). New York, NY: Macmillan.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*(2), 119–142.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology*, *78*(2), 141–146.

Cordray, D., Pion, G., Brandt, C., Molefe, A., & Toby, M. (2012). *The impact of the Measures of Academic Progress (MAP) program on student reading achievement* (NCEE Report No. 2013-4000). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Council of Chief School Officers. (2011). *InTASC Model Core Teaching Standards*. Retrieved from http://www.ccsso.org/resources/publications/intasc_model_core_teaching_standards_2011_ms_word_version.html

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24–28.

Goertz, M. E., Olah, L. N., & Riggan, M. (2009). *Can interim assessments be used for instructional change?* (CPRE Policy Briefs). Philadelphia, PA: Consortium for Policy Research in Education.

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *The American Economic Review*, *100*(2), 250–255.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), 589–612.

Heller, J. I. (2010). *The impact of Math Pathways & Pitfalls on students' mathematics achievement and mathematical language development: A study conducted in schools with high concentrations of Latino/a students and English learners*. San Francisco, CA: WestEd.

Helmke, A., & Schrader, F. W. (1987). Interactional effects of instructional quality and teacher judgment accuracy on achievement. *Teaching and Teacher Education*, *3*(2), 91–98.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2008). *A second follow-up year for measuring how benchmark assessments affect student achievement* (REL 2008-No. 002). Waltham, MA: Regional Educational Laboratory Northeast & Islands.

Hickman, J., Fu, J., & Hill, H. C. (2012). *Technical report: Creation and dissemination of upper-elementary mathematics assessment modules*. Princeton, NJ: Educational Testing Service.

Hiebert, J., & Wearne, D. (1983). *Students' conceptions of decimal numbers*. Newark, DE: University of Delaware.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal of Research in Mathematics Education*, 372–400.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D.L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430–511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. (2012). When rater reliability is not enough: Observational systems and a case for the G-study. *Educational Researcher*, *41*(2), 56–64.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology*, *106*(1), 289.

Hoge, R. D. (1983). Psychometric properties of teacher-judgment measures of pupil aptitudes, classroom behaviors, and achievement levels. *The Journal of Special Education*, *17*(4), 401–429.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297–313.

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education*, 258–288.

Jacobs, V. R., Lamb, L. L., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education*, 169–202.

Jacobs, V. R., Lamb, L. L., Philipp, R. A., & Schappelle, B. P. (2011). Deciding how to respond on the basis of children's understandings. In M. Sherin, V. R. Jacobs, & R. A. Philipp (Eds.), *Mathematics teacher noticing: Seeing through teachers' eyes* (pp. 97–116). New York, NY: Routledge.

Johnson, E. M., & Larsen, S. P. (2012). Teacher listening: The role of knowledge of content and students. *The Journal of Mathematical Behavior*, *31*(1), 117–129.

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615–631.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations, student surveys, and achievement gains*. Seattle, WA: The Measures of Effective Teaching Project, The Bill and Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587–613.

Klieme, E., Bürgermeister, A., Harks, B., Rakoczy, K., Ufniarz, K., Blum, W., & Leiß, D. (2010). *Effects of assessment, grading, and feedback on grade 9 mathematics students*. Unpublished manuscript.

Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *Journal for Research in Mathematics Education*, *37*(4), 297–312.

Konstantopoulos, S., Miller, S. R., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis*, *35*(4), 481–499.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology*, *100*(3), 716.

Martínez, J. F., Stetcher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment*, *14*, 78–102.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572–606.

National Board for Professional Teaching Standards. (1989). *What Teachers Should Know and Be Able to Do*. Retrieved from http://www.nbpts.org/sites/default/files/what_teachers_should_know.pdf

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119.

Peterson, P. L., Carpenter, T., & Fennema, E. (1989). Teachers' knowledge of students' knowledge in mathematics problem solving: Correlational and case analyses. *Journal of Educational Psychology*, *81*(4), 558–569.

Philipp, R. A., Ambrose, R., Lamb, L. L., Sowder, J. T., Schappelle, B. P., Sowder, L., . . .Chauvot, J. (2007). Effects of early field experiences on the mathematical content knowledge and beliefs of prospective elementary school teachers: An experimental study. *Journal for Research in Mathematics Education*, *38*(5), 438–476.

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, *121*(2), 183–212.

Quint, J. C., Sepanik, S., & Smith, J. K. (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessments of Student Thinking in Reading (FAST-R) in Boston elementary schools*. New York, NY: MDRC.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education*, 8–27.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2008). *Can you recognize an effective teacher when you recruit one?* (NBER Working Paper 14485). Cambridge, MA: National Bureau of Economic Research.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*, 537–571.

Rowland, T., Huckstep, P., & Thwaites, A. (2005). Elementary teachers' mathematics subject knowledge: The knowledge quartet and the case of Naomi. *Journal of Mathematics Teacher Education*, *8*(3), 255–281.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, *50*(5), 1020–1049.

Scates, D. E., & Gage, N. L. (1958). Explorations in teachers' perceptions of pupils. *Journal of Teacher Education*, *9*(1), 97–101.

Shavelson, R. J., & Borko, H. (1979). Research on teachers' decisions in planning instruction. *Educational Horizons*, *57*, 183–189.

Shavelson, R. J., & Kurpius, A. (2012). Reflections on learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 13–26). Rotterdam, the Netherlands: Sense Publishers.

Sherin, M., Jacobs, V., & Philipp, R. (Eds.). (2011). *Mathematics teacher noticing: Seeing through teachers' eyes*. New York, NY: Routledge.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–23.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247.

Star, J. R., & Strickland, S. K. (2008). Learning to observe: Using video to improve preservice mathematics teachers' ability to notice. *Journal of Mathematics Teacher Education*, *11*(2), 107–125.

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning*, *10*(4), 313–340.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762.

Toste, J. R., Compton, D. L., Fuchs, D., Fuchs, L. S., Gilbert, J. K., Cho, E., . . . Bouton, B. D. (2014). Understanding unresponsiveness to Tier 2 reading intervention: Exploring the classification and profiles of adequate and inadequate responders in first grade. *Learning Disability Quarterly*, *37*, 192–303.

West, M. R., Morton, B. A., & Herlihy, C. M. (2016). *Achievement Network's investing in innovation expansion: Impacts on educator practice and student achievement*. Cambridge, MA: Center for Education Policy Research.

Wilson, S. M., Shulman, L. S., & Richert, A. E. (1987). "150 different ways of knowing: Representations of knowledge in teaching." In J. Calderhead (Ed.), *Exploring teachers' thinking*. Sussex, UK: Holt, Rinehart, & Winston.

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach For America in high school. *Journal of Policy Analysis and Management*, *30*(3), 447–469.

Manuscript received July 9, 2015
Final revision received November 11, 2017
Accepted February 22, 2018