Teachers' knowledge of students: Defining a domain

Heather Hill, Mark Chin, & David Blazar
Harvard Graduate School of Education

Knowledge of students ranks high among the teacher capabilities identified by 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 effective teaching. Many observe that such knowledge enables a variety of effective classroom strategies, including adjusting the pacing of instruction (Clark & Peterson, 1986), forming appropriate instructional groups (Shavelson & Borko, 1979), and designing tasks and questions to further student understanding (An, Kulm, & Wu, 2008). In fact, several broad-scale interventions and professional development efforts are based on the idea that improving teachers' knowledge of students will improve student outcomes, and evidence exists that some of these approaches have worked (Bell, Wilson, Higgins, & McCoach, 2010; Black & Wiliam, 1998; Carpenter, Fennema, Chiang, Peterson, & Loef, 1989; Tirosh, 2000).

Despite this considerable consensus on the importance of teachers' knowledge of their students, a review of this research literature suggests two quite divergent approaches to measuring this construct. One research tradition, begun in the late 1960s and active for almost two decades, measured teacher accuracy in predicting student performance on cognitive assessments (Helmke & Schrader, 1987; Hoge & Coladarci, 1989; Scates & Gage, 1958). Another research tradition, begun more recently (Carpenter, Fennema, Peterson & Carey, 1988; Hill, Ball & Schilling, 2008) and still active today (Sadler, Sonnert, Coyle, Cook-Smith, & Miller, 2013), focuses on teachers' understanding of student thinking, including students' prior knowledge, likely developmental pathways, and prevalent misconceptions. However, no studies compare these perspectives, describing their relationship to one another and to other aspects of teacher knowledge. Nor do existing studies provide evidence regarding the reliability and validity of the metrics developed to represent these two aspects of teacher knowledge. Finally, few such measures are convincingly related to student outcomes. Without evidence regarding these issues, it is difficult to assess claims that teachers' knowledge of students constitutes a key capability for effective teaching.

In this paper, we describe and evaluate two metrics that tap teachers' knowledge of their students' mathematical thinking. These measures capture teacher *accuracy* in predicting student performance and, following Sadler et al. (2013), teacher *knowledge of student misconceptions*. Specifically, we ask:

1) How well do these metrics differentiate among teachers?
2) Do teachers' scores on these measures show evidence of convergent and discriminant validity?
3) How well, if at all, do scores on these constructs predict student outcomes?

Following our review of prior research, we discuss the data and analyses that led to this conclusion.

Prior Research

Teacher knowledge has long been conceptualized as a multi-faceted construct. Shulman and colleagues' classic formulation of teacher knowledge includes content knowledge, general pedagogical knowledge, pedagogical content knowledge, and knowledge of learners and their characteristics, among other topics (Shulman, 1986, 1987; Wilson, Shulman, & Richert, 1987). In mathematics, the topic of our study, Ball, Thames, and Phelps (2008) elaborated this list, arguing for the presence of two distinct sub-domains within pedagogical content knowledge: knowledge of content and students, and knowledge of content and teaching. Rowland and colleagues (2005) arrived at a different conceptualization, describing foundational knowledge of content (propositional knowledge learned while in the academy); transformational knowledge, as foundational knowledge applied in classrooms; connective knowledge between topics and across lessons; and contingent knowledge, or the ability to reason in practice. Notably, these authors all developed their ideas about these facets from observing teaching practice, albeit of different populations (i.e., novices, in the case of Shulman and Rowland and colleagues.) or through different mediums (i.e., videos of teaching, for Ball and colleagues).

Leaving for others the correct formulation of a broader structure for teacher knowledge, we note the ways in which several theoretical writings provide guidance about a particular hypothesized subdomain and our metric of interest here, teachers' knowledge of students, and in particular the multiple elements that are thought to compose this specific domain. Shulman (1986) writes that knowledge of learners 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 student misconceptions and knowledge of topics that are easy and difficult for students. Ball, Thames, and Phelps (2008) list a more extensive set of sub-domains within this construct, including student conceptions and misconceptions around specific content, student interests, and likely student reaction to particular instructional tasks. Not surprisingly, a recent review (Depaepe, Verschaffel, & Kelchtermans, 2013) suggests considerable consensus in the field around knowledge of student conceptions and misconceptions as composing a major element of pedagogical content knowledge.

Using insights afforded by such theory, scholars have empirically investigated teachers' knowledge of student thinking, including specific student misconceptions, developmental pathways, and problem-solving strategies. We focus in particular on studies that have attempted to create measures of these constructs, a significant portion of the empirical literature (Depaepe et al., 2013). For instance, Carpenter and colleagues (1988) measured teachers' general knowledge of student addition and subtraction strategies as well as the particular strategies deployed by specific students in their classrooms. Bell, Wilson, Higgins, and McCoach (2010) presented teachers with written student work and asked them to identify and explain student errors, comment on a range of student solution strategies for a single problem, and explain what students might have been thinking as they answered. Although Carpenter and colleagues present no reliability information for their items, Bell et al. (2010) used a rubric to score their data, resulting in acceptable estimates of internal consistency reliability (0.76 and 0.79 on the pre-test and post-test, respectively).

In a study more explicitly focused on measurement within this domain, Hill, Ball, and Schilling (2008) used the same ideas about knowledge of students and content (KCS) described in Ball, Thames, and Phelps (2008) to design multiple-choice items. Because the study collected pilot data on a large (teachers, *n*=1552) scale, the authors were able to subject these items to factor analyses for the purpose of construct validation, as well as to tests of 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 (0.58 to 0.69). Teachers' growth from a pre-test to post-test on these items were responsive to their reports of learning about KCS-related topics in a mathematics professional development program, but not to teachers' reports of learning subject matter knowledge, suggesting convergent and discriminant validity. However, with low reliabilities and poor fit to the population—most teachers scored well on these items, producing a ceiling effect—these items were not pursued further (Hill, personal communication).

Despite this healthy interest in measuring teachers' knowledge of student conceptions and misconceptions, few studies related teacher knowledge of these topics to student outcomes. In one exception, Carpenter and colleagues (1988) found no relationship between their measures of teachers' knowledge of student strategies and actual student performance on tests of computation and word problem solving. In a second exception, Sadler, Sonnert, Coyle, Cook-Smith and Miller (2013) constructed a measure of students' science misconceptions by asking teachers to identify the most commonly chosen distractor for each of 20 multiple-choice items. Although as with other studies in this field, the authors did not report the reliability of their assessment, they did 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 only. There was no such effect for low-achieving students.

Concurrent with this substantial interest in knowledge of student misconceptions, a second, seemingly separate, line of work arose from educational psychologists interested in how teachers' knowledge of students might support their judgments during instructional decision-making. Initial empirical work in this field appears to have arisen around estimating teacher accuracy in predicting student performance as an end in itself, documenting teachers' absolute knowledge of their students rather creating usable metrics to predict instruction or student outcomes (see Hoge & Coladarci, 1989). These studies utilized a variety of measurement techniques. For instance, in Coladarci (1986), teachers were asked to anticipate each student's performance on selected achievement test items. The author then differenced the teacher's prediction with students' actual scores for each student-item combination before averaging that difference score to the teacher level. Less detailed (and time-consuming) versions of these metrics asked teachers to predict class rankings or the total scores of their students on an achievement test. Analysts typically then correlated teacher judgments with actual student performance, with a median correlation of 0.66 (Hoge & Coladarci, 1989). Interestingly, significant cross-teacher variability in the accuracy of predictions appeared to exist (Coladarci, 1986; Hoge, 1983; Hoge & Coladarci, 1989).

Two studies used this cross-teacher variability in accuracy to predict student performance. Helmke & Schrader (1987) compared 32 fourth and fifth grade teachers' predictive accuracy in mathematics 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. This measure correlated with students' own reports that the teacher was aware of their performance levels, suggesting concurrent validity, and the metric was marginally significant in predicting students' mathematics test scores (p=0.10), remarkable given the small sample size.

In a study examining the effect of professional development on teachers' knowledge and practice, Carpenter et al. (1988) administered a measure of early-grade teacher accuracy in predicting students' correct and incorrect answers. For each of six randomly selected students, teachers recorded whether each student would successfully solve each of six different addition and subtraction word problems. Teachers were credited for matches between predictions and true student outcomes, and this metric was correlated with students' scores on a computational and word problem assessment. In both cases, correlations were above 0.3; however, no controls were used for prior student achievement (i.e., these were student status, not gain, scores), rendering these results open to critiques of selection bias. In addition, neither the Carpenter nor the Helmke studies presented information on the reliability of their metrics—that is, the extent to which accuracy scores accurately reflected teachers' true level of this trait.

Thus both theory and our review of prior empirical work suggest teacher knowledge is multi-faceted, and that within this particular facet – teachers' knowledge of students – multiple elements may exist. Yet rarely are multiple elements from this domain measured and analyzed in a single study; this means that neither the scores nor the characteristics of scores from the accuracy and student thinking traditions have been compared to one another. Further, only two studies within each of the student thinking and accuracy traditions link teachers' performance to student achievement outcomes, with mixed results in both. Finally, few authors in this field have explored the reliability and validity of the metrics they establish. Below, we describe an effort to address these issues.

Data and Methods

Sample and Setting

For our analyses, we use data collected by the National Center for Teacher Effectiveness (NCTE) main study, which developed and validated different mathematics-related measures in fourth- and fifth-grade classrooms. The NCTE study recruited 583 teachers in four large urban East coast public school districts. Ultimately, 328 of these teachers were deemed eligible to and actually participated in the study. From these teachers and the students in their classrooms, NCTE collected data from the 2010-2011 school year through the 2012-2013 school year. Our analytic sample comprises a subset of the data collected on these teachers and students due to missing data caused by the movement of teachers and classrooms into and out of the study. We describe our restricted sample in more detail below.

Data

Project staff collected data describing teachers, classroom, and students from five sources: teacher questionnaires, video-recorded mathematics lessons, student questionnaires, student performance on a project-developed mathematics test, and district administrative data containing student demographic information and state standardized test performance.

Teachers responded to questionnaires in the fall and spring semesters of each school year, for a possible total of six data collection points. Questions on the fall questionnaire probed teachers' background, preparation, beliefs, and practices in the classroom. A separate set of items assessed teachers' knowledge of the mathematics of upper elementary school and middle school using a mix of items from the Mathematical Knowledge for Teaching instrument (Hill, Rowan, & Ball, 2005) and released items from the Massachusetts Test for Educator Licensure (Charalambous, Hill, McGinn, & Chin, in preparation). Questions on the spring questionnaire probed teachers' coverage of different grade-level mathematical topics in addition to assessing different measures of teachers' knowledge of students, our main construct of interest. We describe these measures in detail below. Teacher survey response rate exceeded 95% for all six data points.

The project aimed to video-record teachers during their mathematics instruction up to three times per school year for each year of the study. Teachers chose when to record lessons, though the study asked teachers not to record lessons on days of testing or test preparation. Lessons averaged approximately one hour in length. In total, the NCTE project video-recorded and scored 1,713 lessons; due to scheduling issues and issues with technology rendering certain videos unusable, teachers averaged just under three video-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 that contains questions designed to elicit students' perception of their mathematics classrooms (Ferguson, 2012). In addition to Tripod questions, the survey also contained items that captured student background characteristics. Across all three academic years, the study collected student questionnaire responses from 94% of the students with project data (i.e., surveys or performance on the project-developed test) in NCTE classrooms.

Participating teachers' students also completed a study-developed mathematics test in the fall and spring semester of each school year (Hickman, Fu, & Hill, 2012). Project staff designed this assessment to include more cognitively challenging and mathematically complex items than those found in many state standardized tests in hopes that it would prove more strongly aligned to the mathematics-specific observational and knowledge measures of teacher effectiveness contained in the study. Across all three academic years, the study collected student performance data from the spring administration of the project-developed test from 95% of the students with project data in NCTE classrooms. Student scores on the test were estimated using a two-parameter logistic item response theory (IRT) model (Hickman et al., 2012).

Finally, each district provided for all fourth- and fifth-grade students the following: mathematics teacher assignment for the 2010-2011 through 2012-2013 school years and up to two years prior to the study; student demographic data, including gender, race or ethnicity, eligibility for subsidized lunch (FRPL), English language learner (ELL) status, and special education (SPED) status; and student test scores on state standardized mathematics and English language arts (ELA) exams.

Measures

Knowledge of student misconceptions and accuracy. As noted above, the teacher questionnaire contained questions intended to assess two different aspects of teachers' knowledge of students. The first measure captured by items on the yearly spring teacher questionnaire was modeled after research investigating teachers' knowledge of student misconceptions (*KOSM*) (Sadler et al., 2013). To measure *KOSM*, the questionnaire used between eight and seventeen selected items from the project-developed mathematics test, asking teachers, "Which [of the following options for this item] will be the most common incorrect answer among fourth [or fifth] graders **in general**?"[1] The NCTE project chose items based on

[1] 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

examination of pilots of the student assessment. First, the project excluded items for which there was not a dominant incorrect student response (i.e., items for which most students answered correctly, and for which incorrect responses were spread equally across distractors). After the project examined each item, researchers prioritized ones for which the dominant student response was well-established in the research literature.

[Insert Figure 1]

[Insert Figure 2]

For instance, the first sample item (Figure 1) probes students' understanding of the equal sign – as an indicator to compute (yielding the most common incorrect answer of 12) or as an indicator that the expressions on either side of the sign are equal to one another (yielding an answer of 7). Because the project-administered test did not contain enough items reflecting common student misconceptions[2], however, we also accepted items that had a simple dominant wrong answer, as in Figure 2, where 43% of fourth graders who answered the question incorrectly chose 24 of the three offered incorrect responses.

To generate *KOSM* scores for each teacher, we first compared teachers' responses to this question to the actual modal incorrect response of fourth or fifth graders.[3] We then estimated the following one-parameter logistic IRT model within grade,[4] using the gsem command in Stata [version 13.1]:

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

In equation (2), y_{it} reflects whether teacher t correctly guessed the modal incorrect response among students for item i of the project-developed mathematics test (i.e., $y_{it} = 1$).[5] Our IRT model controls for α_i, which allows items to have different difficulty parameters. Due to our sample size of teachers and items, we use a one-parameter logistic IRT model, and thus do not vary the discrimination parameters among items, m. From equation (2), we recover each teacher's *KOSM* score, θ_t. To investigate the cross-year stability of this measure of teachers' knowledge of students, we estimate equation (1) within school year and grade.

general in light of low reliabilities observed in the first year of data collection. The study hypothesized that low reliabilities might have resulted from the fewer students taking any particular test item within a classroom, and sought a larger sample size of students from which to judge most common incorrect answer. This question and measure was omitted from the 2012-2013 spring teacher questionnaire.

[2] The goal of the project-administered test was not exclusively to capture common student misconceptions, but instead to measure students' mastery of material contained in Common Core guidelines for each grade. We also perceived that well-documented student misconceptions were relatively rare at these grades at the time the project-administered test was designed (2010).

[3] In 2010-2011, we compared teachers' responses to this question to the actual modal incorrect response of his or her students, if available.

[4] Of the 306 teachers for whom we estimated *KOSM* scores for, nine received scores using items for both fourth and fifth grade. We investigated the effects of removing these teachers from our analyses and find that our results do not substantively change. We report results including these teachers.

[5] For three *KOSM* items from the 2011-12 fourth grade spring questionnaire, and for two *KOSM* items from the 2011-12 fifth grade questionnaire, the most common incorrect response among students represented less than 50% of the responses that students chose incorrectly. When excluding these items from the construction of the *KOSM* measure, however, our results do not change.

The second measure was modeled on research exploring the extent to which teachers can predict student performance on specific material. To measure this construct of teachers' *accuracy*, we used the above items and others from the project-developed mathematics test, this time asking teachers, "Approximately what percentage of **your students** being tested today will choose the correct answer [for this item]?"[6] To generate *accuracy* scores for each teacher, we calculated the actual percentage of correct student answers for each item, differenced each from the teachers' estimate for that item, and then used the absolute values of those differences in the following multilevel equation:

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

The outcome in equation (2) represents this absolute difference for teacher t on item i, rounded to the nearest integer. The model also includes a vector of item fixed effects, α_i, included to capture differences in item difficulty (e.g., higher "difficulties" reflected items where teachers' guesses were further off), and teacher random effects, θ_t, teachers' underlying *accuracy* scores.[7] Because *accuracy* items differed across grades of the study, we estimated equation (2) within grade. Consequently, teachers received *accuracy* scores for each grade in which they responded to the questionnaire.[8] In addition to estimating *accuracy* scores within grade from items across all years of the study, we also estimated equation (1) within school year in addition to grade specifically in order to investigate the cross-year stability of scores and the relationship between year-specific *accuracy* scores to student test performance; we describe these analyses in further detail below. We multiplied all scores for the *accuracy* measure by -1 such that higher scores reflect more accurate predictions.

Other teacher questionnaire measures. Items included on fall questionnaires allowed us to calculate several measures for use in convergent and discriminant validity analyses. We separated the convergent group into two categories: constructs similar to the *accuracy* and *KOSM* metrics, and constructs we expected, based on theory, to correlate with better performance in each of these areas. For similar constructs, we included the expectation that the two constructs would be related to one another, as *accuracy* and *KOSM* both fall under the broad knowledge of students domain, as described above.

Second, we also included teacher's mathematical knowledge for teaching as a theoretically similar construct; teachers' own content knowledge may facilitate their perception of student performance, and stronger knowledge of the mathematics may also help alert teachers to student misconceptions. Based on factor analysis results (see Charalambous et al., in preparation), we constructed a single metric from the Mathematical Knowledge for Teaching (Hill et al., 2005) and Massachusetts Test of Education Licensure (*MKT/MTEL*) assessments, with a marginal reliability of 0.92.

[6] This is the exact wording as it appears on the 2011-2012 and 2012-13 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 substantially similar.

[7] We estimated scores using the same model in equation (1) while including weights to account for differences between teachers in the number of students answering each item. Our results do not change substantively when using *accuracy* scores estimated with weights. We report results using only scores estimated without weights.

[8] Of the 315 teachers for whom we estimated *accuracy* scores for, 17 received scores using items for both fourth and fifth grade. We investigated the effects of removing these teachers from our analyses and find that our results do not substantively change. We report results including these teachers.

In addition to this knowledge metric, we intuited that teaching experience would correlate with *accuracy* and *KOSM* because novice teachers (i.e., those with less than or equal to two years of experience) have had fewer opportunities to observe students' misconceptions than more experienced teachers, and because such teachers face many unfamiliar challenges during their early years of teaching, inhibiting the ability to diagnose student proficiency.

We also hypothesized that teachers' *accuracy* and *KOSM* scores would be related to teachers' grading habits, as more exposure to student work would yield better insight into students' thinking. Teachers' grading habits were measured by a one-item scale asking for an estimate of the number of hours devoted to grading math assignments in a typical week. To arrive at a better estimate of teachers' grading habits (and the other questionnaire measures described below), we leveraged the additional information gained from asking teachers this item across fall questionnaires from multiple years by estimating the following multilevel model:

$$(3)\ y_{yt} = \beta_0 + \alpha_y + \theta_t + \varepsilon_{yt}$$

The outcome of equation (3), y_{yt}, represents teacher t's response to the grading item in year y. The equation includes fixed effects for year, α_y, and random effects for teacher, θ_t. These random effects capture teachers' "underlying" grading habits, or an estimate of their habits when using all available data; furthermore, empirical Bayes shrinks less reliable estimates, due to fewer data points (Raudenbush & Bryk, 2002), closer to the mean.

Our analyses of teachers' use of formative assessment followed a similar logic: these practices are meant to increase teachers' knowledge of their students, and thus should theoretically lead to better scores on our two metrics. We measured formative assessment practices using responses to five items on each fall questionnaire; these questions asked teachers how frequently they evaluated student work using rubrics, provided feedback with numeric scores, differentiated assignments based on student needs, examined the problem solving methods of student work, and asked students to self-evaluate work ($\alpha = 0.59$). After averaging responses across items within a year, we recovered estimates of teachers' underlying formative assessment habits using the same model represented by equation (3).

For discriminant validity analyses, we selected five teacher questionnaire constructs that we expected to be unrelated to teachers' knowledge of students:

- Possession of a master's degree, measured with a single item. Masters' degrees typically do not include coursework that would lead to heightened teacher understanding of students;
- The amount of time and effort teachers report preparing for class. This scale contained teacher responses to three items that asked about activities unrelated to students (i.e., gathering lesson materials, reviewing lesson content, preparation for class by working through explanations; $\alpha = 0.77$);
- Teachers' self-reported classroom climate, measured from eight items including the extent to which students and teachers exhibit care and concern and students comply with behavioral guidelines ($\alpha = 0.89$). We do not expect better climates to lead to better knowledge of student thinking;
- The extent to which teachers report altering the breadth and depth of instruction in reaction to standardized testing, captured by seven items on the fall questionnaire (e.g.,

omitting topics not tested, teaching item formats likely to appear on the state test; $\alpha = 0.85$);

- Teachers' self-reported efficacy, including the extent to which they feel prepared to meet the general challenges that teachers face, measured with 14 total items across years, drawn from work conducted by Tschannan-Moran and colleagues (1998).[9]

Each teacher received a single score on each of these measures, estimated using the method outlined by equation (3).

Video-recorded lessons of mathematics instruction. Video data was scored on both the Mathematical Quality of Instruction (MQI) observational instrument (Hill et al., 2008) and the Classroom Assessment Scoring System (CLASS) observational instrument (Pianta, LaParo, & Hamre, 2007). We followed a similar process – identifying convergent and discriminant constructs – from the measures available to us from each instrument. We hypothesized that teachers' *accuracy* and *KOSM* would be related to two items from the MQI: teacher remediation of student mistakes (*remediation*) and incorporation of student thinking into lessons (*use of student productions*). We similarly hypothesized that items from the CLASS instrument capturing the constructs of *classroom climate* and *classroom organization* would be unrelated to teachers' knowledge of their students.

To obtain teacher-level scores for the two MQI items, each recorded lesson was scored in 7.5-minute segments by two raters who were screened, certified, and trained by the study. Lessons were assigned to raters in order to minimize the number of instances that a rater watched the same teacher. A prior generalizability study demonstrated that observation systems similar to the one used by the NCTE study can yield reliable scores (see Hill, Charalambous, & Kraft, 2012). To arrive at 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, where lessons are nested within teachers:

$$(4)\ y_{lt} = \beta_0 + \theta_t + \varepsilon_{lt}$$

The outcome, y_{lt}, represents teacher t's score for either *remediation* or *use of student productions* on lesson l.[10] The model also contains teacher random effects, θ_t, which also reflects teacher t's underlying MQI score. Using empirical Bayes (Raudenbush & Bryk, 2002), teacher scores are adjusted to account for differences in reliability caused by varying numbers of lesson-level MQI scores included in the model.

To obtain teacher-level scores for the two CLASS dimensions, each recorded lesson was scored in 15-minute segments by a single rater who attended biweekly calibration meetings conducted by the study to help ensure standardization of scoring practices. Similar to the scoring structure of the MQI, lessons were assigned to raters in order to minimize number of instances that a rater watched the same teacher. To arrive at *classroom climate* and *classroom organization*

[9] Between 2010-2011 and 2011-2012, the items assessing teachers' self-reported efficacy were changed on the fall questionnaire. The Cronbach's alpha for 2010-2011 efficacy items was 0.63, and for 2012-2013 efficacy items was 0.86. We standardize average efficacy scores within year before estimating equation (3) to account for the differences in scales.

[10] Due to differences in scoring protocols between observational instruments, 1,694 of the 1,713 video-recorded mathematics lessons were scored with the CLASS, while all were scored with the MQI.

scores for each teacher, we first averaged scores for individual codes of the CLASS across segments and raters to the lesson level. We then took the average across the items within each of the two dimensions. The *classroom climate* dimension comprises scores from nine individual assessments of classroom phenomena, such as the positive climate, teacher sensitivity, quality of feedback, or student engagement ($\alpha = 0.90$); the *classroom organization* dimension comprises scores from codes assessing the negative climate of the classroom (reverse coded), the teachers' behavioral management skills, and the productivity of the classroom ($\alpha = 0.72$). Taking these lesson-level averages for each dimension, we then estimated the same multilevel model depicted in equation (4) to arrive at underlying scores for teachers.

Student questionnaires. Finally, teachers in our analysis received scores based on their students' responses to a questionnaire distributed in the spring semester of each school year. Again, we separated constructs on the student questionnaire into those we expected would be convergent and discriminant with knowledge of student thinking. In the former category, we identified four items relevant to teachers' *accuracy* score: whether students' teachers know when the class does or doesn't understand; whether their teachers actually ask whether the class understands; whether their teachers ask questions to be sure the class is following along during math instruction; and whether their teachers check to make sure that the class understands what is being taught ($\alpha = 0.70$). Students responded to each item on a scale of 1 (Totally Untrue) to 5 (Totally True). To estimate a *monitoring, evaluation, and feedback* score for each teacher from these four student questionnaire items, we first averaged responses across items for each student. We then estimated the following multilevel model, where students are nested within years (i.e., a teacher-year interaction effect), which are nested within teachers:

$$(5)\ y_{syt} = \beta_0 + \theta_t + \nu_{yt} + \varepsilon_{syt}$$

The outcome, y_{syt}, represents the average of the responses to the four monitoring, evaluation, and feedback items of student s in year y taught by teacher t. The model also contains year 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 are adjusted to account for differences in reliability caused by varying numbers of students taught by each teacher.

We conducted the same measure construction process using student responses to items on the questionnaire probing students about the *perceived classroom behavior*. The three items used in the measure, also scaled from 1 to 5, asked students whether time is wasted during instruction (reverse coded), whether students behave badly (reverse coded), and if their classmates behave the way the teacher wants them to ($\alpha = 0.61$).

Analysis Strategy

This paper seeks to understand how well the *accuracy* and *KOSM* measures differentiate among teachers, whether teachers' scores on these measures show evidence of convergent and discriminant validity with other theoretically related and unrelated metrics, and how well teachers' scores on these metrics predict student outcomes. We outline a strategy for answering each question below.

Differentiating among teachers. To ascertain the extent to which these metrics differentiate among teachers and contain information regarding teachers' knowledge of students, we estimate a variety of reliability metrics. For the *accuracy* measure, we estimate the signal-to-noise ratio in teacher scores using estimates of the intraclass correlation (ICC) statistic. For the *KOSM* measure, we use estimates of the marginal reliability produced in the IRT model described above. In addition to estimating the ICC and marginal reliability, we also estimate *accuracy* and *KOSM* scores within school years and examine cross-year correlations, a measure of consistency. When conducting these analyses, we opt to utilize the largest sample of teachers possible (i.e., 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. Results are separated by grade, as the student items used to create these two metrics were specific to the tests at each grade level.

Convergent and discriminant validity of scores. To investigate the convergent and discriminant validity of scores, we correlated teacher *accuracy* and *KOSM* scores with scores from measures constructed using data from teacher questionnaires, video-recorded lessons, and student questionnaires, as outlined above. In these correlational analyses, we combine the sample of fourth and fifth grade teachers, as we have no expectation of different patterns in convergent and discriminant validity scores across grades. For example, we would hypothesize that teachers who spend more time grading student work would have stronger *accuracy* and *KOSM* scores, regardless of the grade level of their students. In addition to this theoretical motivation, combining samples allows for more power in our correlational analyses.

For our analyses investigating the convergent and discriminant validity of scores, and for those investigating the relationship between knowledge of student measures to student outcomes (described below), we restrict our sample to 272 teachers. These teachers have estimates for all of our hypothesized convergent and discriminant metrics, and teach in classrooms that are included in our models predicting student outcomes.

Predicting student outcomes. Most research and policy-makers operated under the assumption that teachers' knowledge of students is predictive of student test gains. We cannot test this hypothesis in a causal manner, as we do not randomly assign students to teacher knowledge levels. However, we examine associations between these forms of teacher knowledge and student outcomes under conditions that attempt to limit the bias in estimates in order to provide suggestive evidence of such a relationship. To do so, we estimate models where teachers' knowledge of student scores predict student performance on either the state standardized mathematics test or the project-developed mathematics test. We estimate our models using the following multilevel equation, where students are nested within years, which are nested within teachers:

$$(6)\ y_{spcgyt} = \beta_0 + \alpha X_{sy-1} + \delta D_{sy} + \phi P_{pcgyt} + \kappa C_{cgyt} + \eta + \omega\theta_{gt} + \psi\pi_t + \mu_t + \nu_{yt} + \varepsilon_{spcgyt}$$

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,[11] in year *y*, taught by teacher *t*. Equation (6) contains the following controls:

- 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, district, and grade-by-year fixed effects;
- θ_{gt}, *accuracy* and *KOSM* scores for teacher *t* for grade g;[12]
- π_t, a vector of other teacher-level variables including teachers' *MKT/MTEL* scores and an indicator for having taught two or fewer years in year y;
- 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.

In addition to restricting the student sample to those taught by the 272 teachers considered in our convergent and discriminant validity analyses of *accuracy* and *KOSM*, a variety of restrictions were placed on the student sample included in this multilevel model, resulting in a sample of 9347 students taught by these teachers. These further restrictions attempted to exclude atypical students (i.e., those who skipped or repeated a grade for either tested outcome) and classrooms (i.e., classrooms with high proportion of special education students, classrooms with high proportion of students with missing baseline scores, and classrooms with fewer than five students).

The multilevel model depicted in equation (6) controls for the student-level predictors used frequently by states, districts, and research studies to obtain value-added scores for teachers. Some predictors incorporated in the model, however, are unique. First, we control for classroom- and cohort-level averages of prior test performance and demographics as well as school fixed effects in our model. If teachers' *accuracy* and *KOSM* scores influence the relationship between these predictors and student test performance, including these controls will attenuate the observed relationship between knowledge of students and outcomes. We opt, however, to choose this restrictive model in order to help address the observational nature of analyses, and address the possibility that students are sorted into teachers' classrooms (see, for example, Rothstein, 2009).

Second, in an effort to disentangle the effect of teachers' knowledge of students from related predictors, our multilevel model contains several teacher-level controls. One control is teachers' mathematical knowledge; as noted above, mathematical knowledge may be related to knowledge of students. We also include a dummy variable that indicating if the teacher has fewer than two or fewer years of experience; several studies (e.g., Kane, Rockoff, & Staiger, 2008)

[11] Fewer than 1% of students took the state standardized mathematics test at a different grade level than the project-developed mathematics test. We consider this percentage to be negligible for the purposes of our analyses.

[12] We investigate the effect of using teacher *accuracy* and *KOSM* scores standardized within grade and district from teacher scores in the model depicted by equation (6). We rescale scores because student state test performance is standardized within district, grade, and school year; this modification does not result in significant differences in our findings.

have shown that novice teachers post slightly weaker value-added scores, and as noted above, experience may play a role in enhancing both *accuracy* and *KOSM*.

In addition to our main investigation into the relationship between teacher *accuracy* and *KOSM* scores to student performance on state standardized mathematics tests and the project-developed mathematics test, we also investigate: the relationship between within-year (as opposed to "career") *accuracy* and *KOSM* scores to student performance, and the relationship between teachers' career knowledge of students scores to student test performance for students at varying levels of prior test performance. We conduct the first of these two additional sets of analyses because these knowledge measures may be composed both of an underlying trait (i.e., knowledge of students) and a year-specific deviation (i.e., knowledge of particular students). Our second analysis explores results found by Sadler and colleagues (2013), who noted heterogeneous effects of teachers' knowledge on outcomes for different groups of students, stratified based on pretest scores.

Results

First, we provide figures depicting teachers' performance on spring questionnaire items assessing their *accuracy* and *KOSM*.

[Insert Figure 3]

[Insert Figure 4]

[Insert Figure 5]

[Insert Figure 6]

The figures above show that the items used to estimate teachers' *accuracy* and *KOSM* scores vary substantially in terms of their difficulty. For the *accuracy* items, the median, among teachers, for the absolute differences between predicted proportion of correct student responses to actual correct range from approximately 10% up to 50%, depending on the grade-level and administration of the spring questionnaire. Similarly, across the *KOSM* items, we see substantial amounts of variation in terms of the proportion of teachers correctly guessing the modal incorrect response for students on different items of the project-administered mathematics test.

Differentiating among Teachers with *Accuracy* and *KOSM* Scores

To investigate how well the *accuracy* measure differentiated among teachers in our sample, we estimated adjusted ICCs. Specifically, we calculated the proportion of the variance in y_{it} of equation (1) attributable to differences between teachers, after controlling for item fixed effects. The unadjusted ICC reflects the signal-to-noise ratio of teacher performance on a single item measuring *accuracy*; because each teachers' score comprises their performance on multiple accuracy items from the spring questionnaire(s), we adjust the ICC for the total number of accuracy items that the average teacher responded to $(n = 20)$. We find the adjusted ICC of *accuracy* scores for fourth grade teachers to be 0.74, and for fifth-grade teachers to be 0.71. 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 0.79 and 0.78, respectively. These values for the adjusted ICC suggest that, with enough item responses, our measure of *accuracy* can somewhat reliably differentiate among teachers' performance on this construct.

To investigate how well the *KOSM* measure differentiated among teachers in our sample, we calculated marginal reliability statistics following the estimation of our IRT model reflected in equation (2). 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). The estimated marginal reliability of *KOSM* scores for fourth grade teachers was 0.21; for fifth grade teachers it was 0.40. The magnitude of the average standard errors of scores reflects these reliability coefficients, suggesting very imprecisely estimated scores for a given individual; for fourth grade, the average magnitude was 0.93, and for fifth grade, the average magnitude was 0.85. The low estimates of reliability and the noisiness of score estimates suggest that the *KOSM* measure does not adequately differentiate teachers.

The difference in reliability between the two metrics can also be seen when comparing the cross-year correlation of within-year scores. Depending on the grade and combination of years, the cross-year correlation of *accuracy* score is moderate, ranging from 0.32 to 0.51; these estimates suggests that teachers' abilities to predict the proficiency of their students is somewhat consistent from school year to school year, despite changes in the population of students being instructed. Furthermore, the range of correlations compares favorably to the cross-year correlations of other measures of teacher quality used by policymakers and researchers (Goldhaber & Hansen, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2008; Polikoff, 2015). *KOSM* scores, on the other hand, demonstrate less consistency from one year to the next. For fourth grade teachers, scores correlated at 0.22 between 2010-2011 and 2011-12. For fifth grade teachers, this correlation was slightly higher at 0.26.

The differences between *accuracy* and *KOSM* measures of teachers' knowledge of students might emerge for several reasons. Because the *accuracy* questions were asked on three different administrations of the spring survey, and the *KOSM* questions on just two, the overall reliability, operationalized through the adjusted ICC or marginal reliability metrics, should be higher for the former construct. The cross-year correlations for the *KOSM* measure may be lower because of changes to the language of questions assessing the construct between the 2010-2011 version and the 2011-2012 version of the spring questionnaire.

Convergent and Discriminant Validity of *Accuracy* and *KOSM* Scores

Next, we investigate the direction and magnitude of correlations between teacher *accuracy* and *KOSM* scores to other measures that we hypothesize should and should not be related to teachers' knowledge of students. First, we consider the results of our convergent validity analyses, seen in Table 1.

[Insert Table 1]

To start, we observe an extremely low and insignificant relationship between our two constructs of interest, *accuracy* and *KOSM*. This may owe to the poor reliability of the *KOSM* measure, but it is still surprising, given that both are theorized to be nested within the same general domain of pedagogical content knowledge. We do observe, as predicted, a positive and significant relationship between both constructs and teachers' mathematical knowledge as measured by the *MKT/MTEL* variable, supporting the interpretation that knowledge of teaching-related mathematical content and students are related, as hypothesized in the pedagogical content knowledge literature.

Table 1 shows few significant relationships between the experiences thought to enhance *accuracy* and *KOSM* and teachers' scores on these metrics. Being a novice teacher is associated with lower scores on both measures, but not significantly so. The amount of reported time grading student work had a mixed relationship to the two metrics, and the relationship for formative assessment was nearly significant, but in the opposite direction as expected, with teachers who report more formative assessment practices performing less well on these two measures.

Correlations between *accuracy* and *KOSM* with raters' and students' indicators of closely related teaching activities appeared somewhat more promising. Teachers who were observed as offering more classroom remediation did post modestly higher scores on both metrics, though the interpretation of these correlations differs by construct. For the first, teachers who can better predict their own students' likely performance also offer slightly more remediation, regardless of whether that students' performance is strong or weak. This suggests that performance on both metrics may be governed by teachers' general awareness of students. For the second construct, teachers with stronger *KOSM* also engaged in more in-class remediation during observed lessons. The directionality is not clear in this case – whether *KOSM* develops from teacher engagement in remediation, or whether stronger *KOSM* spurs more remediation activity. *Accuracy* also appears strongly related to teachers' use of student productions, suggesting that teachers who incorporate more student ideas into their lessons can also better predict their students' performance. A similar relationship did not appear for student productions and *KOSM*. Finally, teacher performance on *accuracy* and *KOSM* was largely not related to students' reports of teacher monitoring, evaluation, and feedback, although this may be reflective of the low reliability of the latter metric ($\alpha = 0.59$).

[Insert Table 2]

We now turn to discriminant validity. From Table 2, we see that relationships between these constructs and teachers' *accuracy* and *KOSM* were, as hypothesized, non-existent. There are two exceptions. First, possession of a master's degree is positively and marginally significantly related to the two constructs of interest. However, this may stem from the fact that experience, theoretically a contributor to accuracy, is associated with further education in our data. We test this possibility by running partial correlations between master's degree and accuracy, controlling for whether the teacher is a novice, and find that the relationship between the first two measures is indeed weakened when accounting for the experience of teachers. The other exception to our non-correlations is in the area of effort, where teachers who performed worse on our *KOSM* and *accuracy* measures also reported more time spent on preparing for class. We can think of no reason for this correlation, and note that our expectation for violation of divergent validity would have been a positive effect—that is, higher-scoring individuals tend to perform better on diverse sets of metrics.

Overall, results from these analyses are quite mixed. Although correlations with constructs hypothesized to be divergent are quite low, correlations between some constructs hypothesized to be convergent were also quite low. One reason may be the generally poor reliability for some of these metrics, and perhaps poor validity for some of the self-report data. Nevertheless, we take the modest correlations observed in Table 1, on convergent validity, as evidence that at least some of the structures and relationships we hypothesized are supported. We turn now to predicting student outcomes.

Predicting Student Outcomes

As noted above, both theory and prior empirical evidence suggests that *accuracy* and *KOSM* scores should be related to student test performance. The results of our explorations can be seen in Table 3 below.

[Insert Table 3]

From Table 3, we see that teachers' *accuracy* scores strongly associate with 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 and demographic characteristics, or school fixed effects. The magnitude of the effect of teachers' *accuracy* scores is fairly large. As seen in Table 3, the standard deviation (SD) of teacher random effects is approximately 0.11 (i.e., students taught by a teacher one-SD above average in terms of effects on student performance on the project-developed test, on average, score 0.11 SDs higher). The point estimate on teacher *accuracy* thus suggests that being taught by a teacher one-SD above average, in terms of *accuracy*, approximates the effect of being taught by a teacher who is 0.5 SDs above average, in terms of effects on student outcomes. Alternatively, we can compare the magnitude of the coefficient to the effect of being a disadvantaged student, socioeconomically, on performance on the project-developed test; specifically being eligible for a free- or reduced-price lunch is associated with a 0.034 SD decrease in test performance (not shown). Conversely, teachers' scores on the *KOSM* measure showed no relationship to their students' performance.

We also investigate the association of the two knowledge of students measures to 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 constructs. This raises a question of whether teachers' *accuracy* and *KOSM* scores predict student test performance more generally, or whether we have identified a test-specific phenomena. Yet, as Table 3 shows, though we see slightly weaker relationships overall between *accuracy* scores and student test performance on the state standardized mathematics tests, the direction of the effect is consistent with our findings regarding student performance on the project-developed mathematics test; we observe a suggestive positive relationship (i.e., $p < 0.11$). The consistent positive relationship between teachers' *accuracy* scores and their students' test performance on different tests corroborates hypotheses posited by prior literature.

Similar to the project-administered test, we observe no association of teachers' *KOSM* scores to their students' state test performance. Earlier investigation into how well the *KOSM* measure differentiated among teachers suggested that scores on this second measure of teachers' knowledge of students were fairly unreliable. Despite this fact, standard errors for the point estimates of the effect of *KOSM* on student test performance were comparable to those for the point estimates of the effect of *accuracy* on student test performance.

[Insert Table 4]

Table 4 above further supports our conclusions regarding student outcomes. In this table, we use within-year knowledge of student scores in order to explore the possibility that knowledge of student constructs may measure a year-specific component of teacher capability as well as a general trait. Again, we find that teacher *accuracy* scores positively predict student test performance on both the project-developed test and the state standardized test, with stronger associations demonstrated with performance on the project-developed test.

[Insert Table 5]

Finally, Table 5 above shows the results from our exploration into a finding reported by Sadler and colleagues (2013). Similar to Sadler et al., we find that the effect of teachers' knowledge of students (in our case, our significant predictor, *accuracy*) on student outcomes differs across student populations. Specifically, we find suggestive evidence for a weaker relationship between *accuracy* and test performance for students who are lower performing at baseline. We are at a loss to explain this.

Conclusions

This paper provides an example of how one might investigate novel metrics for capturing teachers' knowledge of students. Although other efforts in this area have faltered due to measurement challenges (Hill et al., 2008), and still others have returned mixed results with regards to the relationship of such knowledge to student test performance (Carpenter et al., 1988; Helmke & Schrader, 1987), we believe the results here are promising. While these results are not causal, we controlled for student sorting by using a fairly rigorous model (i.e., one incorporating peer- and cohort-level averages of student demographics and baseline test performance in addition to school fixed effects) on the student outcome side; we also find it less-than-plausible that student achievement could cause teacher familiarity with student work, though in non-experimental research, anything is possible.

Our investigation produced fairly good evidence that teachers' *accuracy* in predicting students' outcomes is an identifiable teacher-level trait. This metric returned adequate reliabilities for both fourth and fifth grade teachers and teachers' scores were correlated across years, suggesting that teachers are fairly consistent in their ability. This metric was also related to teachers' mathematical knowledge and their engagement with specific classroom activities we hypothesized would be related to accuracy, in particular remediation of student misconceptions and use of student productions. This metric also predicted student outcomes on the project-administered test from which it was derived, and was marginally significant in models predicting student performance on state standardized tests. These mild relationships and significant predictive power may be interpreted in several ways. If our evidence of construct validity is correct and further developed in future analyses, we might consider accuracy to be a separate and important facet of teachers' knowledge of students. However, we also recognize that our evidence is indeterminate on this point; accuracy may be simply an indicator of more generally high-quality instructional practice. Either way, these results suggest that a construct roughly titled "knowing where your students are, in terms of mastery of content" should take a place in contemporary theoretical delineations of teacher knowledge

The story is more complicated for *KOSM*, where our fourth and fifth grade metrics performed considerably more poorly. Reliabilities were sub-par, although there was some evidence of cross-year correlation in teacher scores and weak evidence for convergent validity. However, this metric did not predict student outcomes. We find this somewhat surprising, given the central place of *KOSM* in most theories of teacher knowledge (Ball et al., 2008; Shulman, 1986), although we also note that the two empirical studies of this area returned mixed results (Carpenter et al., 1988; Sadler et al., 2013). This suggests that analysts may turn attention to alternative measurement strategies within this domain. The development of knowledge on student learning trajectories, for instance, poses one avenue for inquiry.

Finally, we don't have a good measure of what teachers actually do with such knowledge. Although we opened this article with evidence from observational and case studies of teaching, we argue that more can and should be learned about how knowledge in this arena supports student learning. Strong teacher accuracy may result in better matching of mathematics lessons to students' skill level; more appropriate amounts of time spent remediating students' mathematical misunderstandings, or other by-products. These seem like issues that could benefit from further exploration in both observational and experimental settings.

Works Cited

An, S., Kulm, G., & Wu, Z. (2004). The pedagogical content knowledge of middle school, mathematics teachers in China and the US. *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.

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*, 479-512.

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

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., 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.

Charalambous, C. Y., Hill, H. C., McGinn, D., & Chin, M. (in preparation). *Teacher knowledge and student learning: Bringing together two different conceptualizations of teacher knowledge*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

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: 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.

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

Depaepe, F., Verschaffel, L., & Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research. *Teaching and Teacher Education*, *34*, 12-25.

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

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.

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.

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

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, *39*(4), 372–400.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008b). Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study. *Cognition and Instruction*, *26*, 430-511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When Rater Reliability is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, *41*, 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.

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.

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.

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

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

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2007). *Classroom Assessment Scoring System (CLASS) Manual.* Baltimore, MD: Brookes Publishing.

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

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

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.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational researcher*, 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.

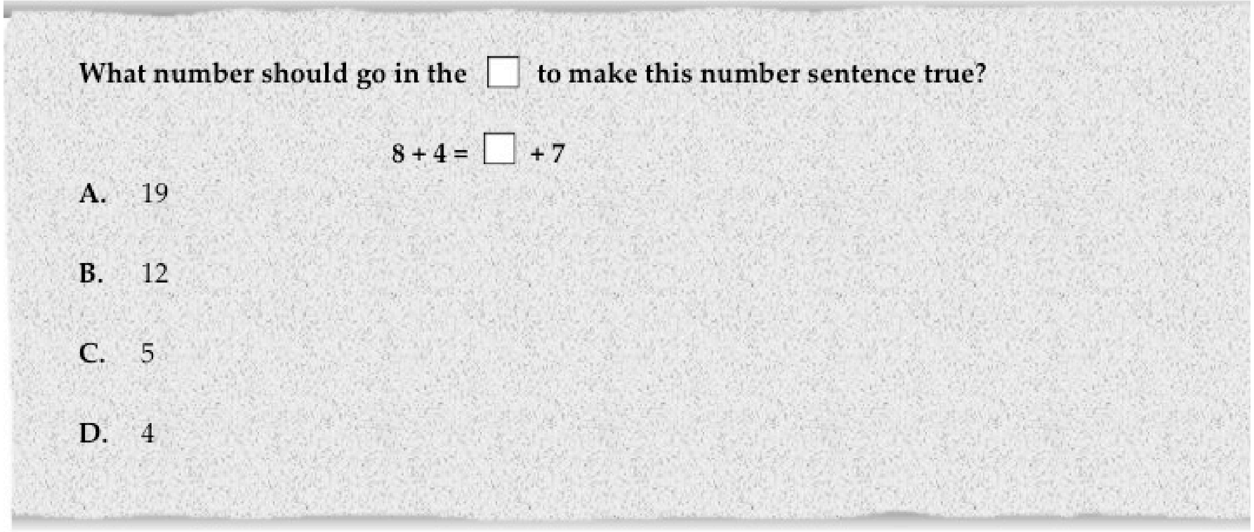
Tirosh, D. (2000). Enhancing prospective teachers' knowledge of children's conceptions: The case of division of fractions. *Journal for Research in Mathematics Education*, 5-25.

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher Efficacy: Its Meaning and Measure. *Review of Educational Research*, *68*, 202-248.

Wilson, S. M., Shulman, L. S., & Richert, A. E. (1987). "150 different ways" of knowing: Representations of knowledge in teaching.

Figures and Tables

Consider the following problem from the student assessment:



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

$$8 + 4 = \square + 7$$

A. 19

B. 12

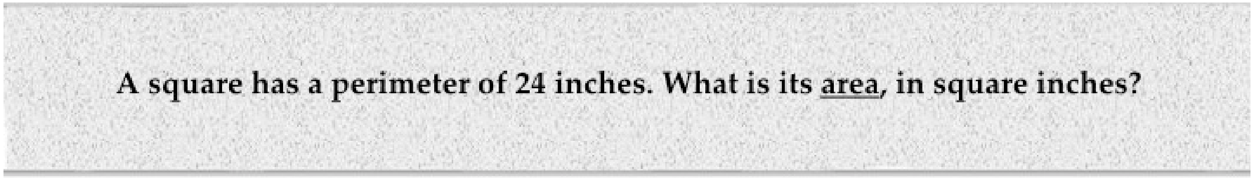
C. 5

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 1. Example item on spring questionnaire used to assess both *accuracy* and *KOSM*. For the *KOSM* measure, this item has a researched-aligned dominant incorrect student response.

Consider the following problem from the student assessment:



A square has a perimeter of 24 inches. What is its area, in square inches?

The correct answer to this problem is 36. a. Approximately what percentage of **your students** being tested today will answer correctly?	%
b. Approximately what percentage of fourth grade students in **your district** will answer correctly?	%
Which will be the most common incorrect answer among fourth graders **in general**? (Please circle **ONE** answer.)	6 24 96

Figure 2. Example item on spring questionnaire used to assess both *accuracy* and *KOSM*. For the *KOSM* measure, this item has a simple dominant incorrect student response.

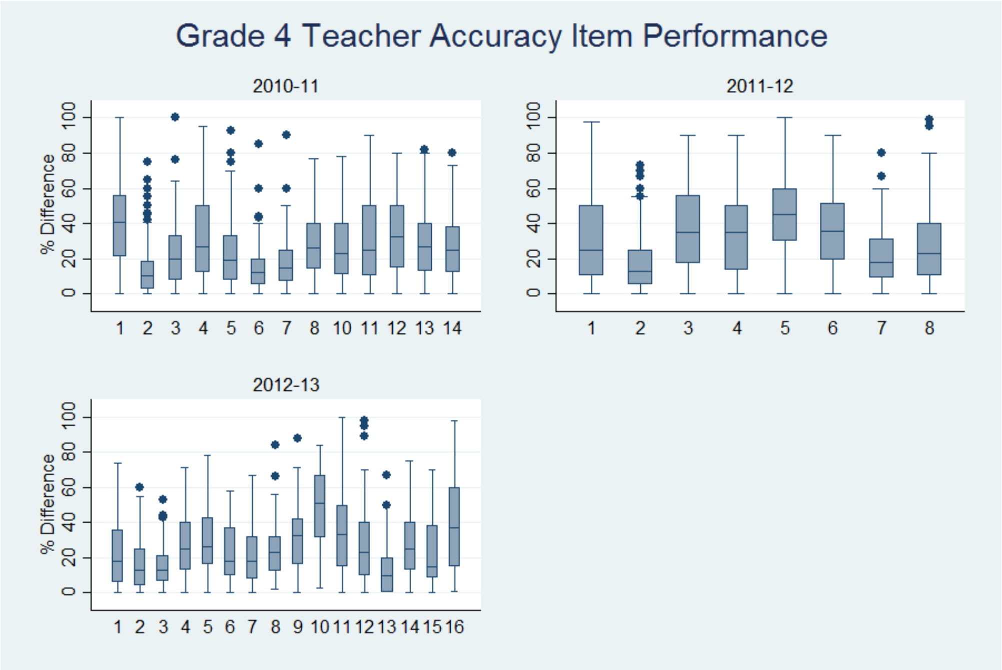


Figure 3. Teacher performance on *accuracy* items in fourth grade.

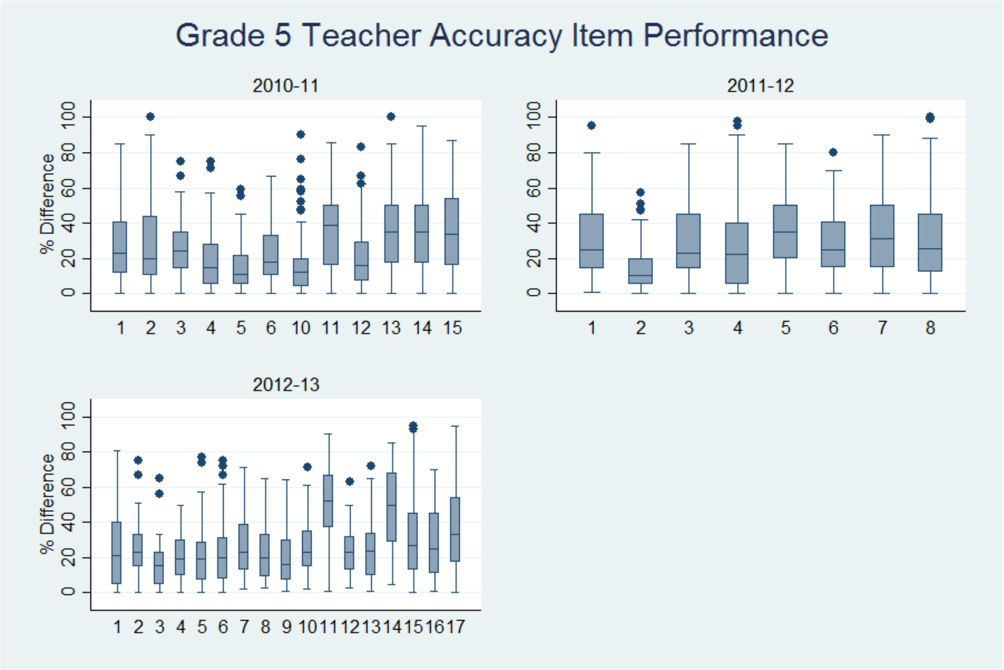


Figure 4. Teacher performance on *accuracy* items in fifth grade.

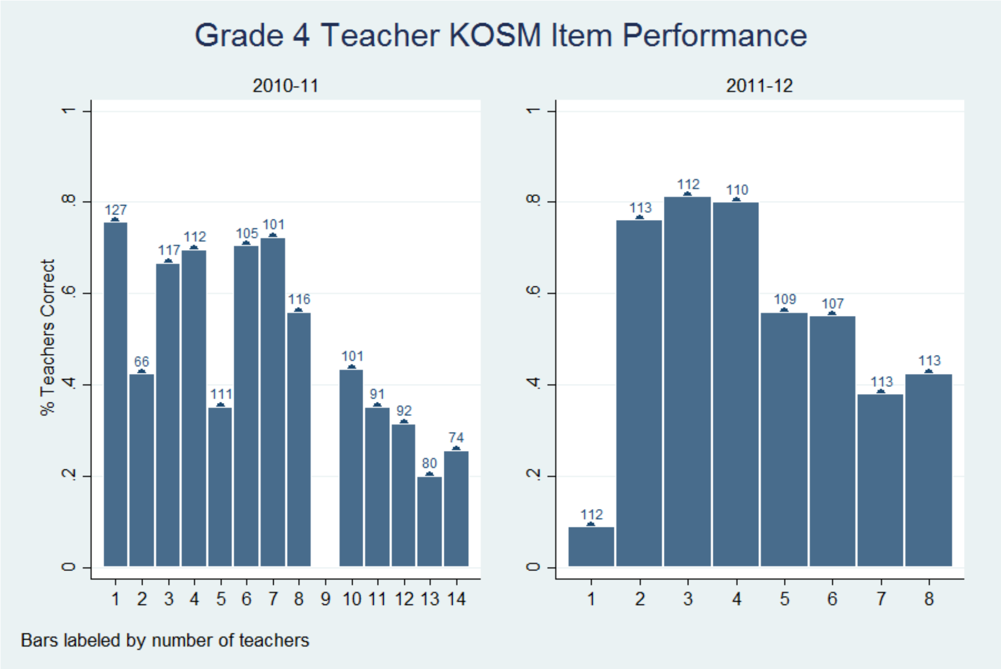


Figure 5. Teacher performance on *KOSM* items in fourth grade.

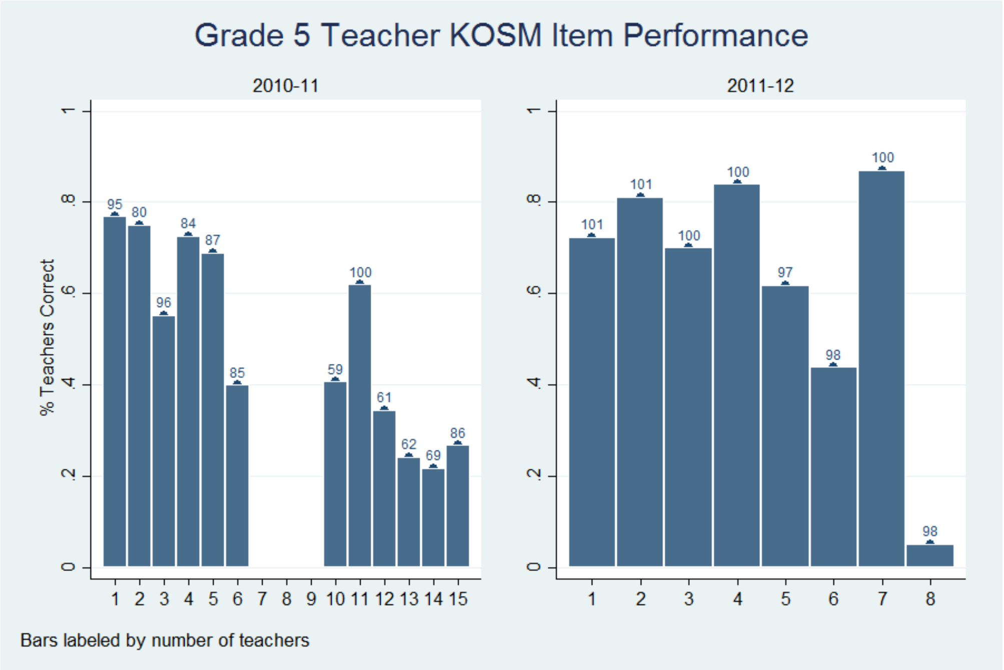


Figure 6. Teacher performance on *KOSM* items in fifth grade.

Please do not cite or distribute without consent of the authors.

Table 1. Convergent Validity of Knowledge of Student Scores

Measure	Accuracy	KOSM	MKT/MTEL	Novice (<=2 Years)	Grading Habits	Formative Assessment	Remediation	Use of Student Productions	Monitoring, Evaluation, and Feedback
Grades 4 & 5									
Accuracy	1	0.08	0.25***	-0.04	0.05	-0.11~	0.14*	0.22***	0.01
KOSM	0.08	1	0.13*	-0.08	-0.08	-0.08	0.12*	-0.03	-0.11~

Note: Number of teachers is 272. ~p<0.10, *p<0.05, **p<0.01, ***p<0.001

Please do not cite or distribute without consent of the authors.

Table 2. Discriminant Validity of Knowledge of Student Scores

Measure	Master's Degree	Effort	Self-Reported Classroom Climate	Test Prep	Efficacy	Classroom Climate	Classroom Organization	Perceived Classroom Behavior
Grades 4 & 5								
Accuracy	0.10~	-0.17**	-0.02	0.10	-0.04	-0.02	-0.02	0.09
KOSM	0.10	-0.16**	-0.08	0.02	-0.06	-0.05	-0.02	-0.09

Table 3. Predicting Student Test Performance with *Career* Knowledge of Student Scores

	Student Mathematics Test Performance					
	Project-Developed Test			State Standardized Test		
Teacher Knowledge of Student Predictors						
Accuracy	0.049***		0.049***	0.024		0.024
	(0.013)		(0.013)	(0.015)		(0.015)
Knowledge of Misconceptions		-0.014	-0.015		0.011	0.011
		(0.012)	(0.012)		(0.014)	(0.014)
Other Teacher Predictors						
Novice Teacher (<= 2 Years)	-0.045	-0.042	-0.046	-0.131*	-0.129*	-0.130*
	(0.051)	(0.052)	(0.051)	(0.054)	(0.054)	(0.054)
Career MKT/MTEL	0.018	0.028*	0.020	0.024	0.027~	0.022
	(0.013)	(0.013)	(0.013)	(0.015)	(0.015)	(0.015)
Value-added Model Predictors						
Student Prior Test Performance Vector	x	x	x	x	x	x
Grade-Year Interaction Fixed Effects	x	x	x	x	x	x
Student Demographic Vector	x	x	x	x	x	x
Classroom-level Aggregates	x	x	x	x	x	x
Cohort-level Aggregates	x	x	x	x	x	x
School Fixed Effects	x	x	x	x	x	x
District Fixed Effects	x	x	x	x	x	x
Teacher Random Effects	x	x	x	x	x	x
Teacher-Year Interaction Random Effects	x	x	x	x	x	x
SD of Teacher Random Effects	0.110	0.116	0.109	0.148	0.150	0.148
	(0.016)	(0.017)	(0.016)	(0.015)	(0.015)	(0.015)
Number of Students	9347	9347	9347	9347	9347	9347
Number of Teachers	272	272	272	272	272	272

Table 4. Predicting Student Test Performance with *Within-Year Knowledge* of Student Scores

	Student Mathematics Test Performance					
	Project-Developed Test			State Standardized Test		
Teacher Knowledge of Student Predictors						
Accuracy	0.054***		0.054***	0.024*		0.023~
	(0.011)		(0.013)	(0.012)		(0.014)
Knowledge of Misconceptions		-0.009	-0.011		0.019	0.018
		(0.013)	(0.013)		(0.015)	(0.014)
Other Teacher Predictors						
				-		
Novice Teacher (<= 2 Years)	-0.042	-0.056	-0.058	0.128*	-0.120*	-0.119*
	(0.051)	(0.055)	(0.053)	(0.053)	(0.059)	(0.058)
Career MKT/MTEL	0.019	0.037*	0.028*	0.024~	0.030~	0.028~
	(0.013)	(0.015)	(0.014)	(0.015)	(0.016)	(0.016)
Value-added Model Predictors						
Student Prior Test Performance Vector	x	x	x	x	x	x
Grade-Year Interaction Fixed Effects	x	x	x	x	x	x
Student Demographic Vector	x	x	x	x	x	x
Classroom-level Aggregates	x	x	x	x	x	x
Cohort-level Aggregates	x	x	x	x	x	x
School Fixed Effects	x	x	x	x	x	x
District Fixed Effects	x	x	x	x	x	x
Teacher Random Effects	x	x	x	x	x	x
Teacher-Year Interaction Random Effects	x	x	x	x	x	x
SD of Teacher Random Effects	0.113	0.112	0.101	0.142	0.148	0.140
	(0.017)	(0.021)	(0.023)	(0.015)	(0.018)	(0.019)
Number of Students	9256	7633	7587	9256	7633	7587
Number of Teachers	272	272	272	272	272	272

Table 5. Predicting Student Test Performance with *Career* Accuracy Scores, Interactions

	Student Mathematics Test Performance			
	Project-Developed Test		State Standardized Test	
Teacher Knowledge of Student Predictors				
Accuracy	0.049***	0.074***	0.024	0.038*
	(0.013)	(0.017)	(0.015)	(0.017)
Interactions				
Accuracy * Prior Test Performance	0.001		0.013~	
	(0.008)		(0.007)	
Accuracy * 1st Tercile Prior Test Performance		-0.044**		-0.033*
		(0.017)		(0.014)
Accuracy * 2nd Tercile Prior Test Performance		omitted		omitted
		(.)		(.)
Accuracy * 3rd Tercile Prior Test Performance		-0.027		-0.003
		(0.018)		(0.015)
Other Teacher Predictors				
Novice Teacher (<= 2 Years)	-0.045		-0.130*	
	(0.051)		(0.054)	
Career MKT/MTEL	0.018		0.024	
	(0.013)		(0.015)	
Value-added Model Predictors				
Student Prior Test Performance Vector	x	x	x	x
Grade-Year Interaction Fixed Effects	x	x	x	x
Student Demographic Vector	x	x	x	x
Classroom-level Aggregates	x	x	x	x
Cohort-level Aggregates	x	x	x	x
School Fixed Effects	x	x	x	x
District Fixed Effects	x	x	x	x
Teacher Random Effects	x	x	x	x
Teacher-Year Interaction Random Effects	x	x	x	x
SD of Teacher Random Effects	0.110	0.111	0.148	0.147
	(0.016)	(0.016)	(0.015)	(0.015)
Number of Students	9347	9347	9347	9347
Number of Teachers	272	272	272	272