STATE OF NEW MEXICO
LEGISLATIVE EDUCATION STUDY COMMITTEE

REPRESENTATIVES
Rick Miera, Chair
Nora Espinoza
Mary Helen Garcia
Jimmie C. Hall
Dennis J. Roch
Mimi Stewart

ADVISORY
Alonzo Baldonado
Ray Begaye
Eleanor Chávez
George Dodge, Jr.
Roberto "Bobby" J. Gonzales
Tim D. Lewis
Sheryl M. Williams Stapleton
Shirley A. Tyler
Bob Wooley

State Capitol North, 325 Don Gaspar, Suite 200
Santa Fe, New Mexico 87501
Phone: (505) 986-4591 Fax: (505) 986-4338
http://www.nmlegis.gov/lcs/lesc/lescdefault.aspx

SENATORS
Cynthia Nava, Vice Chair
Mary Jane M. García
Gay G. Kernan
Lynda M. Lovejoy

ADVISORY
Vernon D. Asbill
Mark Boitano
Stephen H. Fischmann
Howie C. Morales
John Pinto
Sander Rue

Frances Ramírez-Maestas, Director
David Harrell, PhD, Deputy Director

December 14, 2011

MEMORANDUM

TO: Legislative Education Study Committee

FR: Michael Bowers

RE: CONTROLLED RESEARCH ON TEACHER VALUE-ADDED SCORES IN NEW MEXICO

During a presentation on Efforts to Improve Teacher Effectiveness at the November 2011 interim meeting of the Legislative Education Study Committee, Dr. Richard Howell, Dean of the University of New Mexico College of Education, offered to provide the committee with a controlled research study on Value-Added Modeling done in New Mexico.

The 2010 article, entitled *"A Validity Argument Approach to Evaluating Teacher Value-Added Scores,"* is attached. The article:

- focuses on the extent to which value-added scores correspond to other indicators of teacher and teaching quality;
- compares the value-added scores of 24 middle school mathematics teachers to survey- and observation-based indicators of teacher quality, instruction, and student characteristics;
- found that teachers' value-added scores correlated not only with their mathematical knowledge and quality of instruction but also with the population of students they teach; and
- illustrates through case studies problems that might arise in using value-added scores in pay-for-performance plans.

American Educational Research Journal
Month XXXX, Vol. XX, No. X, pp. 1–38
DOI: 10.3102/0002831210387916
© 2010 AERA. http://aerj.aera.net

A Validity Argument Approach to Evaluating Teacher Value-Added Scores

Heather C. Hill
Harvard Graduate School of Education
Laura Kapitula
Calvin College
Kristin Umland
University of New Mexico

*Value-added models have become popular in research and pay-for-performance plans. While scholars have focused attention on some aspects of their validity (e.g., scoring procedures), others have received less scrutiny. This article focuses on the extent to which value-added scores correspond to other indicators of teacher and teaching quality. The authors compared 24 middle school mathematics teachers' value-added scores, derived from a large (*N *= 222) district data set, to survey- and observation-based indicators of teacher quality, instruction, and student characteristics. This analysis found teachers' value-added scores correlated not only with their mathematical knowledge and quality of instruction but also with the population of students they teach. Case studies illustrate problems that might arise in using value-added scores in pay-for-performance plans.*

KEYWORDS: accountability, validity, reliability, teacher assessment, educational policy

---

HEATHER C. HILL is an associate professor, Harvard Graduate School of Education, Gutman Library #445, 6 Appian Way, Cambridge, MA 02138; e-mail: *heather_hill@harvard.edu*. Her research interests include teacher quality, mathematics instruction, and teacher mathematical knowledge.

LAURA KAPITULA is an assistant professor, Department of Mathematics and Statistics, Calvin College; e-mail: *lk24@calvin.edu*. Her research interests include quantitative evaluation, value-added modeling, and influence diagnostics.

KRISTIN UMLAND is an associate professor, Department of Mathematics and Statistics, University of New Mexico; e-mail: *umland@math.unm.edu*. Her research interests include teacher mathematical knowledge, teacher professional development, and value-added modeling.

Educational "value-added models" have garnered significant attention in the past decade. These models, which estimate teacher and school effectiveness based on student gains, have become popular in research, evaluation, and pay-for-performance plans, including Minneapolis' Teacher Advancement Program, Dallas's Value-Added Assessment System, Washington, DC's, IMPACT program, and the recently enacted federal Race to the Top competition. Such models and performance plans have proven popular because research indicates that teachers have a large and lasting impact on student achievement (Gordon, Kane, & Staiger, 2006; Rowan, Correnti, & Miller, 2002; Wright, Horn, & Sanders, 1997), because rewarding educators based on effectiveness is thought to motivate better performance (Hanushek, 2007; Schacter & Thum, 2004), and because even critics judge value-added models as more appropriate than cross-sectional models in determining teacher and school efficacy (Amrein-Beardsley, 2008; Linn, 2008; McCaffrey, Koretz, Lockwood, & Hamilton, 2003; Weingarten, 2007).

But what, exactly, do teachers' value-added scores represent? To some, these scores are the most direct indicators of teacher quality and effectiveness. Student learning is the primary goal of schooling, and value-added scores are the most logical, cost-effective method for identifying teachers' contribution to learning (e.g., Gordon et al., 2006; Hanushek, 2007; Hanushek, Kain, O'Brien, & Rivkin, 2005; Wright et al., 1997). Many advocates, in fact, treat value-added scores as if they were objective measures of teacher quality (Duncan, 2009; "Editorial: The New Haven Model," 2009; Hanushek et al., 2005; Weerasinghe, 2008). Other scholars doubt the accuracy and validity of value-added scores, noting that they represent not only some "true" value teachers add to student learning but also the effects of prior teachers, measurement error, and potentially even bias resulting from the distribution of students into classrooms and teachers into schools (e.g., Amrein-Beardsley, 2008; Kupermintz, 2003; McCaffrey et al., 2003). Taken together, these critiques suggest that value-added scores may fail to accurately represent teacher quality.

Despite these critiques and more general concern about the utility of value-added models to measure teacher effectiveness, most research on value-added scores has, to date, been purely quantitative. Missing are studies that examine the relationship between value-added scores and the characteristics they are assumed to represent: good teaching and, by extension, good teachers. To shed light on this relationship, this article describes a mixed-methods study linking teacher value-added scores to their mathematical quality of instruction and a key teacher characteristic, mathematical knowledge for teaching. We also examine relationships between teacher value-added scores and the student characteristics many hope are unrelated to those scores: student background, special education status, and similar attributes. To organize this inquiry, we adopt a validity argument approach

(M. Kane, 2001, 2004). We describe the background, method, and results from this study below.

Measuring Educational Processes and Outcomes

Value-added models, first popularized by Sanders and colleagues (Sanders & Rivers, 1996; Sanders, Saxton, & Horn, 1997; Wright et al., 1997), make use of current and historical test scores to estimate a teacher's effect on student achievement growth. Part of the appeal of value-added scores is based on evidence that teachers exert considerable influence on their students' achievement. In one well-designed study, teacher effects explained 11% of the variation in student test score gains (Nye, Konstantopoulos, & Hedges, 2004; Rockoff, 2004), and in another, teachers' prior-year value-added scores constituted the strongest predictor of future teacher value-added performance (Gordon et al., 2006).

Armed with this evidence, policymakers have adopted value-added techniques in hopes of improving student achievement. Notably, Race to the Top asked states to open the door to value-added accountability systems, both by removing roadblocks to linking teachers and students and by encouraging teacher evaluation plans that include value-added measures as a component. Many states complied (Dillon, 2009, 2010). Second, value-added-based accountability and pay systems are already in wide use in practice. Florida famously pursued value-added-based performance pay with mixed success through the 1990s and early 2000s; Oklahoma and Colorado have both recently adopted pay-for-performance plans that include a value-added component ("Colorado Teacher-Evaluation Bill Enacted," 2010); 16 other states have largely voluntary district-initiated programs. Dallas has long maintained a value-added-based teacher evaluation and pay system, and Houston, Austin, and Washington, DC, have more recently followed suit (see Center for Educator Compensation Reform [CECR], 2010; Lewin, 2010). In fact, of the 65 member districts of the Council of Great City Schools, nearly one fourth have implemented some form of value-added-based school or teacher rewards program.

Concurrent with the rush to adopt value-added models for teacher evaluation and pay, scholars have begun to voice doubts about the accuracy and validity of value-added scores. This literature is quite broad and growing daily. To focus our review, we concentrate on two areas relevant to the work described below: debates about how to properly produce teacher value-added scores and concerns about the validity and reliability of those scores.

On the first point, there is considerable debate about the most appropriate specification for value-added models. One issue is whether to control for student-level covariates in the models. Some argue that teacher scores are stable regardless of the inclusion of these student covariates because the

inclusion of prior-year test scores accounts adequately for student characteristics and allows students to serve as their own controls (Ballou, Sanders, & Wright, 2004). Others argue that failing to adjust for covariates may be unfair to teachers of at-risk students and advocate for models that control for student- and classroom-level factors (Amrein-Beardsley, 2008; Kupermintz, 2003). Another point of debate is how to construct peer groups: whether teachers should be compared to others in the district, others in the same school, others in the same grade within school, or some combination of the above. This corresponds, in a statistical sense, to whether to include school and grade fixed effects in the estimation of teacher-level value-added scores. In practice, it is difficult to accurately disentangle the effects of school, teacher, and grade (McCaffrey et al., 2003). Finally, some scholars now recommend the use of multiple years of data to inform teacher value-added scores (Koedel & Betts, in press).

A survey of districts that generate and use teacher value-added scores for low- or high-stakes accountability shows little consensus around model specification. Dallas, for instance, uses a three-stage model-fitting process that controls for many student demographic and school-level variables and students' previous-year test scores using covariate adjustment (Weerasinghe, 2008). In Denver, the ProComp system uses both conditional growth quantiles and the multivariate model of Lockwood, McCaffrey, Mariano, and Setodji (2007) with student-level covariates to estimate teacher effects (Wiley, Spindler, & Subert, 2010). In the early phase of its teacher accountability system, New York City used covariate adjustment models with grade-, student-, classroom-, and school-level covariates; school fixed effects were not included, and models with both 2 and 3 years of prior student scores were used (Rockoff, Staiger, Kane, & Taylor, 2010). The Education Value Added Assessment System (EVAAS), the most widely used commercially available system, controls for neither student nor school effects but does include district fixed effects in its state-level models. In at least one state, Florida, model details have been left up to districts, with the possibility that a simple gain score, rather than ranks produced by a value-added model, might be used. And in 2009–2010, the first year of its IMPACT program, Washington, DC, had only 1 year of prior student test score data, limiting analyses to covariate adjustment models. Results from these models were recently used, in part, to dismiss more than 200 teachers (Lewin, 2010).

Importantly, while some scholars have recently argued that estimates from multiple years are superior to single-year estimates of teacher effects, it is unclear whether states and districts have heeded that advice or have the data to construct such models. Koedel and Betts (in press) conclude that while using multiple years of data may improve the quality of teacher scores, "often implicitly, the value-added discussion in research and policy revolves around single-year estimates of teacher effects" (p. 4). They also

note that for a large fraction of their data, including novice teachers, multiple years of data were not available (p. 23).

On the second point, major criticisms of value-added models center on the reliability and validity of teacher scores. In fact, investigations into value-added scores have returned relatively low reliabilities. For example, Koedel and Betts (2007) found that although teachers have substantial influence over student outcomes in San Diego Public Schools, variance decomposition also shows only modest reliability, on the order of .57 in mathematics and .46 in reading. Generally in this data set and elsewhere, teachers' value-added scores are composed of roughly equivalent amounts of "error" and "true score" variance (Hanushek et al., 2005; T. J. Kane, Rockoff, & Staiger, 2006; Lockwood, Louis, & McCaffrey, 2002; McCaffrey, Sass, Lockwood, & Mihaly, 2009). Furthermore, two recent studies have shown that teacher value-added scores often vary considerably by the student assessment form and subtests used to construct them (Lockwood, McCaffrey, Hamilton, et al., 2007; Papay, in press).

Results of investigations into the validity of value-added scores are more mixed. Schacter and Thum (2004) find substantively significant correlations between teachers' value-added scores and observational measures of their teaching performance, on the order of .55 to .70. By contrast, elements of a commonly used observational system, Classroom Assessment Scoring System (CLASS; Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008), predicted students' growth trajectories from first through fifth grade only modestly; the authors conclude that the most consistently significant factor, socioemotional qualities of interactions, "matter somewhat" when predicting student growth (Pianta et al., 2008, p. 388). Other studies, including those that compare administrators' ratings of teachers to value-added outcomes, return correlations of between .20 and .50 (Jacob & Lefgren, 2005; Kimball, White, & Milanowski, 2004; Medley & Coker, 1987; Milanowski, 2004).

Results from these inquiries thus leave several unanswered questions. To begin, how can scores possess an arguably unacceptable person-level reliability yet moderate evidence for convergent validity? A chief candidate explanation, according to many critics, is that the apparently strong convergent validity results from spurious correlations due to unmeasured student characteristics. A test for divergence between these characteristics and value-added scores would provide evidence on this point; however, evidence for convergent and discriminant validity has never been examined within a single study.

A related question centers on how high validity correlations must be to support claims of convergent validity. While many take correlations of roughly .60 as strong evidence for convergent validity, M. Kane (2006) notes there are no rules of thumb in this regard and that the degree of acceptable convergence should be determined by the planned use of data. We argue that the target level of agreement between value-added and observational

scores should be based on potential uses of the scores and may in fact be greater than this benchmark. As well, there are no studies that explore the potential consequences of using value-added scores for making specific decisions about specific teachers. We argue that given the widespread use of value-added scores, such studies are urgently needed.

Finally, more research is needed on what value-added scores represent. Many assume that they represent good teaching—and, by extension, good teachers—but observational research that critically examines these assumptions is scarce. In fact, the literature on value-added scores has been almost purely quantitative and studies that complement this literature are also urgently needed.

To structure such a study, we turn to validity theory. In recent years, measurement experts have recommended and illustrated frameworks for inquiry into the validity of many types of assessment (American Educational Research Association/American Psychological Association [AERA/APA], 1999; M. Kane, 2001, 2006; Messick, 1988, 1989). We argue, like others (Amrein-Beardsley, 2008; Kupermintz, 2003), that teachers' value-added scores constitute a form of assessment—one with job-related stakes attached—and thus formal inquiry into their validity should test the appropriateness of inferences and actions based on these scores. To do so, we follow M. Kane's (2001, 2004) argument-based approach. As Kane suggests, we explicitly state assumptions regarding the meaning of teacher value-added scores, then test these assumptions using empirical evidence. Specifically, we focus on the relationship between value-added scores and the processes that are assumed to shape and not shape them as well as the potential consequences of using these scores to identify particular groups of teachers. To frame the assumptions, we considered both actual and proposed high-stakes uses of value-added scores in accountability systems (CECR, 2007; Gordon et al., 2006; Hanushek, 2007; Lewin, 2010; Wright et al., 1997) and standards proposed for educational measurement (AERA/APA, 1999). Our assumptions—and related inferences for empirical testing—are the following:

1. Value-added scores derive from the influence of teacher characteristics and teaching quality on student performance. Thus, value-added scores should correlate more strongly with other indicators of teacher and teaching quality than with hypothetically unrelated constructs. Specifically, value-added scores should,

    a. Converge with expert ratings of instructional quality
    b. Converge with estimates of teachers' knowledge
    c. Fail to correlate with unrelated constructs, such as the population of students in a teachers' classroom

2. In order to affect educational improvement, the use of value-added scores in accountability decisions must not create negative systemwide consequences

for those who populate that system. For this to occur, decisions based on scores must,

a. Identify both excellent and poor teachers with a reasonable degree of accuracy
b. Not distort incentives for educators working within the system

Inferences a, b, and c of the first assumption focus on the extent to which teachers' value-added scores converge with related measures and fail to correlate with theoretically unrelated constructs, often called convergent or discriminant validity. Underlying these inferences is an assumption based on the model of teaching and learning presented in Cohen, Raudenbush, and Ball (2003), Ball and Forzani (2007), and Grubb (2008). This model represents teaching as a set of interactions among teachers, students, and content (the last often seen as instantiated in materials). Teachers, students, and materials hold the primary resources that result in student learning; it is in their interaction during instruction that such learning develops. While more distal factors (e.g., monetary resources, policies) can shape teaching as well, this study focuses on relationships among measures of the central features of the model: teachers' intellectual resources, the instructional behaviors that develop as a product of such resources, and the student learning that results. If value-added scores do represent teacher quality, as many argue, then these three indicators should converge.

Identifying unrelated constructs is more difficult. On its face, Cohen et al.'s (2003) model of instruction suggests that teachers' value-added scores will be related to the resources that students bring to the classroom. Teachers who work with students who are not native English speakers, who have learning disabilities, or who lack access to out-of-school learning opportunities are more likely, in this model, to produce less absolute student growth. In fact, significant debate among value-added researchers has centered on the extent to which students' prior test scores control for the effects of these characteristics (see Ballou et al., 2004; Kupermintz, 2003; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Tekwe et al., 2004). Most agree, however, that for scores to be used in accountability systems, two teachers who bring the same skill to teaching should be identified as equal by the measure regardless of the students they teach. We adopt this standard for our inquiry into discriminant validity.

Our second assumption holds that decisions based on scores should create no negative systemwide consequences. This assumption has two concrete inferences. First, the relationship among teacher quality, teaching, and value-added scores should be sufficiently strong as to accurately categorize almost all teachers during any decision process; any miscategorization could be perceived as unfair to other teachers and school staff. Second, decisions made during the implementation of a value-added-based accountability

system should not distort incentives for teachers, including incentives to serve diverse student populations. To illuminate these inferences, we use case studies to illustrate the consequences of rewarding teachers based on value-added scores.

We conduct our study in the subject of mathematics, where research in the past two decades has been directed toward the measurement of key teacher and instructional characteristics. These efforts allow us to determine the congruence of survey-based, observation-based, and outcome-based measures of teacher quality.

Method

This study included collection of extensive observational, interview, and survey data for a small set of purposively sampled middle school mathematics teachers (n = 24). Following this data collection, we calculated value-added scores for all middle school teachers within the district (N = 222) and extracted the scores for the 24 focal teachers. We then compared value-added and other indicators of teacher quality. Restricting the in-depth sample to a small number of cases implies that this study is limited in several regards. Correlations in small samples are imprecise, and this study thus does not have the power, for instance, to definitively test for differences between convergent and divergent correlation strengths. Nevertheless, the small sample size also imparts several advantages. First, we wanted multiple sources of in-depth data on each teacher in the sample in order to accurately characterize teachers' knowledge and practice; inaccurate characterizations (e.g., low reliability) would leave us open to the possibility that any lack of relationship was due to measurement error. Yet accuracy is expensive; the high reliabilities described below were arrived at through extensive data collection—six lessons per teacher, two 60-minute surveys, and 3 hours of interviews. While more cursory data collection would have allowed a larger sample size, we argue that a carefully executed study is of equal value and provides information that other studies cannot.

Second, this study sought to detect substantively large correlations—and substantively large differences in correlations, in the case of comparisons that examine convergent or divergent validity—rather than to detect weak relationships or differences. Given this, a small sample size provides adequate statistical power. Finally, a primary goal of this research is to provide policy-relevant information in a timely manner; given the widespread adoption of value-added scores, it is incumbent on researchers to inform policymakers' use of those scores carefully but expediently.

District Context and Sampling

This study was conducted in a midsized district in the southwestern United States containing 26 middle schools ranging in size from fewer

than 400 students to more than 1,100 students. Student socioeconomic status was moderate for a large urban district, with 57% of middle school students receiving free or reduced-price lunch (FRL). The district is also racially diverse, with a middle school student population that is 57% Hispanic, 32% Caucasian, 5% American Indian, 4% African American, and 2% Asian. In the years before the study, only four middle schools had more than 50% of their students testing at a proficient level or better in mathematics according to state standards, and, as elsewhere, many schools were failing to make adequate yearly progress under No Child Left Behind. This district was not implementing a value-added-based accountability system for either its teachers or schools at the time of this study.

To identify potential research sites, we fit a series of covariate-adjusted linear mixed models separately to the districtwide data for the 2004–2005, 2005–2006, and 2006–2007 academic year test cycles.[1] The initial models used the current year state mathematics assessment (SMA) score as an outcome and a polynomial function of the previous year's SMA score as well as a series of indicator variables that represented student FRL status, grade, race/ethnicity, and gender. The school effect was entered into the model as either fixed or random, and teacher was generally included as a random effect.[2] Consistent with Tekwe et al. (2004), sensitivity analyses showed that there were high correlations among the rankings (greater than .96) regardless of whether or not school effects were treated as fixed or estimated with empirical Bayes shrinkage and whether or not the teacher random effect was included in the model. Incorporating student-level demographic variables in the model did change teacher ranks. To be conservative, we included student variables in the models that helped identify schools for recruitment. Using these models, we selected six schools for recruitment, prioritizing those with diverse and stable value-added scores and similar demographic descriptors. Ultimately four schools chose to participate: two with high value-added scores, one with moderate scores, and one with low scores. We expected that variation in school value-added scores would increase the likelihood of variation in teacher and instructional quality. Table 1 provides descriptive statistics for each school.

We recruited 26 teachers within these four schools, and 24 agreed to participate. We elected not to recruit most special education teachers and those who taught only a small number of students. One teacher who participated was replaced by a long-term substitute approximately 4 weeks before the start of state testing. This teacher had taught the class for more than 5 months, and because of both this and the fact that we had complete data, we retained her in the models described below. Another teacher was primarily responsible for another subject and taught only one mathematics class. We included her in the analysis but ensured she was not an outlier.

In terms of descriptive information, participants in this study were similar to those in a national sample (see Hill, 2007). The average number of years

Table 1
School Descriptive Statistics

School	*n*	Sp. Ed. (%)	FRL (%)	Avg. MKT Percentile	Avg. Lesson MKT	Avg. MQI	Mean SMA 06-07	Avg. Raw Gain	Percentile Rank 07-08
Barnes	404	12	100	15	1.56	1.79	643	25	92
Gutierrez	468	8	60	41	1.52	1.73	655	18	15
Montgomery	686	9	45	78	1.99	2.39	669	23	81
Watkins	498	11	45	43	1.93	1.93	670	20	50

Note. All school names are pseudonyms. *n* = number of students used to calculate school's rank within district; Sp. Ed. = percentage of students who were in special education; FRL = percentage of students who qualified for free or reduced-price lunch; Avg. MKT percentile = average normalized mathematical knowledge for teaching (MKT) ranking against the national sample reported as a percentile rank; Avg. Lesson MKT = average teacher MKT estimate based on observations; Avg. MQI = average teacher mathematical quality of instruction (MQI) score; Mean SMA 06-07 = mean standardized mathematics assessment student score in the 2006–2007 school year; Avg. Raw Gain = the average gain on the standardized assessment from the previous year; Percentile Rank = percentile score for the school value-added estimate for 2007–2008, using a student background adjusted model with a school fixed effect and teacher random effect.

experience was 12.5, with 6 individuals in their first 5 years of teaching. Teachers' career experience varied, with many working only in middle schools (12), some reporting experience in elementary schools (7), and some in high schools (5). Half (12) reported possessing an undergraduate or graduate mathematics major or minor, and 15 had mathematics-specific credentials. Some specific characteristics of the teachers in this sample are noteworthy. In one school, instruction occurred in Spanish in three classrooms.[3] Teachers tended to be tracked by grade and student ability; for instance, several taught only gifted and talented students, and many taught only one grade.

Data Collection

Teacher data were collected between January and March 2008. Student achievement data were obtained from the district for the 2007–2008 school year in the fall of 2008, and historical achievement data had been obtained in previous years. The study collected six lessons from each teacher, lessons that were scheduled based on simple criteria (no testing days, no field trips, "regular" instruction rather than special lessons designed specifically for the study) and teacher convenience. Most teachers' lessons were videotaped in two waves, with three lessons collected in January and three in late March; because teachers were in a new unit during the second time period, this ensured variation in the content taught.

After each taped lesson, teachers responded to a list of debriefing questions about lesson content, execution, and what students learned.

Videotapes were transcribed, with an estimated 95% of teacher and 50% of student utterances captured audibly.

Teachers also participated in two 60-minute interviews that asked them to describe their practice and their views of mathematics, math teaching, and the students they currently teach. These interviews included a "clinical interview" similar to those used in Hill et al. (2008) that asked teachers to solve and discuss items designed to gauge mathematical knowledge for teaching. Finally, teachers completed two surveys assessing background characteristics and mathematical knowledge for teaching (MKT) with items similar to those used in Hill, Rowan, and Ball (2005). The MKT survey was chosen because past evidence has shown that teacher performance on this instrument predicts student outcomes. Teachers were paid $300 for general study participation and $50 for each returned survey. We had nearly complete participation in all study components; one teacher failed to return the second survey. In her case, we treated items on the second questionnaire as missing; item response theory (IRT) methods are robust to missing data.

Instruments

Student achievement outcomes were based on the SMA, a test given to students in Grades 3–8. The assessment is composed of a mix of multiple-choice (71%), open-ended, and short-answer items drawn from the SAT-10 and Harcourt item pools. An inspection of released items shows they assess a mix of procedural and conceptual knowledge. Form and interrater reliabilities are close to or greater than .90, and vertical equating between adjacent-year tests implies student gain scores can be calculated from adjacent-year test data. The SMA is administered each March, roughly 8 weeks prior to the end of the school year. Because all but 8 weeks of each school year can be assigned to a single teacher, this mitigates problems associated with attributing growth to specific teachers.

MKT was measured by 159 survey items that tap middle school number, operations, and proportional reasoning. MKT goes beyond the typical knowledge a mathematically competent adult may have, focusing on the mathematical knowledge that is specialized to teaching. For example, such knowledge includes providing grade-level-appropriate but precise mathematical definitions, interpreting and/or predicting student errors, and representing mathematical ideas and procedures in ways learners can grasp. Development of this instrument was described in depth in Hill (2007), and the theoretical basis for this instrument is detailed in Ball, Thames, and Phelps (2008).

Finally, we coded lessons using an observational instrument focused on the mathematical quality of instruction (MQI; Hill et al., 2008). The MQI captures the disciplinary integrity of the mathematics presented to students, including the degree to which teacher errors occur during the mathematics

class; it also captures other salient features of mathematics, such as how explanations, representations, precise language, and mathematical generalizations are developed in class. We elected to use this instrument over others because a mathematics-specific instrument was warranted for a study of student mathematics outcomes and because alternative mathematics-specific instruments focused heavily on the degree to which instruction matched "reform" ideals. Because we conducted the study in schools with diverse mathematics curricula, the study required an instrument that was more agnostic with regard to teaching method.

The MQI instrument provides ratings on multiple subscales, such as teacher errors, richness of the mathematics, and student cognitive demand. In this article, however, we focus only on two more general ratings. One is the overall assessment of each lesson's mathematical quality (overall MQI). This 3-point Likert-style rating was applied by trained observers, with *low* corresponding to lessons with significant teacher mathematical errors, *medium* corresponding to lessons with few such errors yet mostly routine instruction, and *high* reserved for lessons with few errors as well as significant mathematical richness through explanation, representations, and strong interactions with students. Interrater agreement of more than 80% was obtained before coding proceeded, and a generalizability study showed that the teacher-level reliability of the MQI rating is .90. Each lesson was individually coded by two raters, then the pair reconciled disagreements. Raters were randomly assigned to one another and to lessons.

The other measure is raters' lesson-based evaluation of teachers' MKT (lesson-based MKT), also coded on a 3-point Likert-type scale using the same procedure outlined above. This estimate differed from overall lesson quality in two ways. First, observers could consider evidence that the teacher was mathematically knowledgeable beyond what was observed in the majority of the lesson. An example would be a lesson that was mostly focused on procedural practice but during which the teacher briefly demonstrated remarkable command of both mathematical explanations and student thinking about mathematics. The lesson would be assigned a high lesson-based MKT score and average MQI score, on the theory that the teacher was expert in the content but the lesson delivered on that day did not consistently require that expertise. Second, in assigning lesson-based MKT ratings, observers could ignore what we thought of as "measurement error" associated with the particular day we sampled instruction. An example would be a lesson with 20 minutes of very strong mathematics instruction and 25 minutes of orientation to high school mathematics coursework. This lesson would receive an average score for MQI, but the lesson-based MKT score would be high. Although this measure was not included in the generalizability study, interrater agreement of 80% was reached before coding proceeded.

Analysis

We averaged teachers' MQI scores over the six lessons coded from videotaped data, then did the same for lesson-based MKT. We obtained an estimate of teachers' written MKT by entering their survey responses into an IRT model and scoring these data alongside those obtained as part of a nationally representative sample of middle school teachers (see Hill, 2007). Two-parameter IRT models were used to score teacher responses with a resulting reliability of .97. We report teachers' scores as the percentile ranks they would have obtained had they taken the MKT instrument as part of the larger sample.

We constructed value-added scores for all middle school mathematics teachers in the district with complete data for nine or more students during the year we studied instruction (N = 222).[4] Students were excluded from these models if they spent less than a full year at the same school, had evidence of switching mathematics teachers midyear, and/or had missing values for the previous year's standardized math, reading, or science scores. Because there is significant disagreement in the literature regarding the most appropriate specification of value-added models (see McCaffrey et al., 2003), eight different models similar to those used in the school selection portion of the study were initially fit to the data. All models explored were covariate adjustment models with minimally all three test scores in the previous year (math, reading, and science) used as covariates. We explored adding test scores from an additional prior year as well, but the estimated teacher effects were highly correlated, r = .98, and to limit the amount of missing data, we used only the previous year's scores. Although covariate adjustment models are known to be biased due to measurement error, Sanders (2006) argues, based on simulation and empirical data results, that when at least three previous test scores are available, as they are here, this is no longer a significant problem.

Ideally, multivariate response models, such as EVAAS, would have been used to estimate value-added scores as they yield estimates with better properties than covariate-adjusted models (McCaffrey et al., 2003; Sanders, 2006). Multivariate response models directly model the entire vector of student scores jointly, but fitting these models requires data for all years and subjects linking students to teachers. These data were not available, a problem that is not unique to this district (see Buddin & Zamarro, 2008). Scores generated from models using teacher fixed effects and empirical Bayes estimates of teacher effects were correlated strongly (r = .98–.99), thus we discuss only empirical Bayes estimates here. Because models with student background variables and school fixed effects did result in different rankings, we present one of each type.

Model 1: Simple model. Adjusting for prior student scores, teacher random effect, the simple model is,

$$y_{ijkl} = \beta_0 + \beta \mathbf{x}_{ijkl} + \tau_{kl} + \xi_{jkl} + \varepsilon_{ijkl},$$

where y_{ijkl} is SMA scale score for the ith student in the jth classroom of the kth teacher at the lth school in 2007–2008 and $\mathbf{x}_{ijkl}$ is a vector consisting of the $ijkl$'s student's previous-year SMA scale score (SMA06), SMA06 squared, SMA06 cubed, grade indicator variables, previous year's standardized reading scale score (SRA06), previous year's standardized science scale score (SSA06) and SMA06, SRA06, and SSA06 by grade indicators interaction terms. β is a parameter vector. The $\tau_{kl} \sim N(0, \sigma_\tau^2)$ are teacher random effects, and ξ_{jkl} and ε_{ijkl} are classroom- and student-level error terms assumed to be independent where,

$$\varepsilon_{ijkl} \sim N(0, \sigma_\varepsilon^2)$$
$$\xi_{jkl} \sim N(0, \sigma_\xi^2)$$

Model 2: School fixed effects model. Adjusting for prior student scores, teacher random effect (as with the simple model), plus a school fixed effect, the school fixed effects model is,

$$y_{ijkl} = \beta_0 + \beta \mathbf{x}_{ijkl} + \tau_{kl} + \varphi \mathbf{c}_l + \xi_{jkl} + \varepsilon_{ijkl},$$

where $\mathbf{c}_l$ is a vector of indicator variables for each school and the ϕ are the parameters indicating the school fixed effects.

Model 3: Student background adjusted model. Adjusting for prior student scores and a teacher random effect (as with the simple model) plus student background variables, the student background adjusted model is,

$$y_{ijkl} = \beta_0 + \beta \mathbf{x}_{ijkl} + \tau_{kl} + \lambda \mathbf{s}_{ijkl} + \xi_{jkl} + \varepsilon_{ijkl},$$

where $\mathbf{s}_{ijkl}$ is a vector of student covariates including indicator variables for accelerated or enriched, algebra, FRL, English language learner (ELL), special education (SPED), Spanish test language, indicator variables for each ethnicity and a SPED by SMA06 interaction, and λ is a parameter vector.

Model 1, the simple model, is similar to the univariate response model of Sanders and Wright (2008). Although Sanders and Wright do not specifically indicate adding higher order polynomial terms in the previous year's test scores, the relationship was curvilinear for these data, and thus these higher order terms were included in the model. Models 2 and 3 represent ideas from more general debates about how best to create value-added scores, including whether to include student background variables and school fixed effects.

Table 2
Correlations Among Teacher Value-Added Scores From Different Models

Model	Simple	School Fixed Effect	Student Background Adjusted	Average Raw Gain
Simple	1	.90	.93	.60
School fixed effect	.90	1	.82	.57
Student background adjusted	.93	.82	1	.57

Note. Spearman rank order correlations are reported (N = 222). Average Raw Gain = the average student gain on the standardized assessment from the previous year.

Table 2 provides the Spearman rank order correlation matrix between scores from the different models. As expected, all three models have fairly high correlations but with some variability. The correlation between the simple model and the student background adjusted model is .93, and the correlation between the simple model and the school fixed effects model is .90. This latter model may attribute teacher effects to their school, while the former may attribute school effects to teachers. As is noted in McCaffrey et al. (2003), school and teacher effects can be difficult to disentangle.

We also present a simple measure of raw student gains, calculated by finding a gain score for each student by differencing SMA07-08 and SMA06-07 and taking the average of those gains for each teacher. This information is similar to that returned to districts by the state of Florida for use in local accountability systems (CECR, 2007).

Study participants' (n = 24) value-added scores were extracted from this data set (N = 222). Each participant has a score for each model as well as for the observational and survey data described above. We related these scores to the MQI and lesson-based MKT measures described above using a Spearman rank order correlation. We did not correct for measurement error in these correlations because our research questions focus, in part, on what kinds of inferences can be drawn using standard (uncorrected) value-added scores.

Results

We begin with basic descriptive information about the data, then organize our results by the assumptions and inferences presented in the introduction to this article.

Data Descriptors

After adjusting for students' previous-year test scores and grade level, we found that districtwide 81% of the variance was at the student level, 4% was at the class level, 13% was at the teacher level, and 1% was at the school

level. Our estimates of teacher-level variance are similar to others in the field (e.g., Nye et al., 2004; Rockoff, 2004). Notably, little of the variance lies between classes within teacher, suggesting that middle school mathematics teachers in this district are rather equally effective over the classes they teach each year. Teacher 2006–2007 and 2007–2008 value-added scores are correlated at .58 to .67, similar to other reports on the year-to-year stability in value-added estimates.

Table 3 shows descriptive data on our sample of teachers. Looking at the value-added rankings according to the simple model (Model 1), only about one third of teachers in our observational sample scored below the 50th percentile as compared to all teachers in the district. This can be explained by the fact that two of our participating schools were from the top quartile. In addition, special education teachers in this district tended to have lower value-added scores than general education teachers, but they were not typically included in the observation sample. The school fixed effects model (Model 2) compares teachers within each school; here the teachers' rank order remains the same within schools, but teachers are more evenly distributed across percentiles. Finally, the student background adjusted model (Model 3) produces similar results to the simple model with the exception of only a few teachers.

Comparisons of teachers within schools suggests that schools differ with regard to their MKT and observational scores. An ANOVA helped formally evaluate differences among the teachers at the four schools in the study. No significant differences were found in teachers' average overall lesson quality within school, $F(3, 20) = 1.82$, $p = .17$, and differences were approaching statistical significance for lesson-based MKT, $F(3, 20) = 2.98$, $p = .06$. However, there were differences in survey-based MKT, $F(3, 20) = 4.23$, $p = .02$. Specifically, higher quality teachers and teaching tended to occur in the more affluent schools. These findings align with other reports (Boyd, Lankford, Loeb, & Wycoff, 2003; Clotfelter, Ladd, & Vigdor, 2004; Hill, 2007; Hill et al., 2005) that find a modest relationship between teacher quality and student characteristics. These findings also suggest that it may be difficult to disentangle the effect of teacher quality and student characteristics on teachers' value-added scores, a point to which we return later.

Finally, as expected given Cohen et al. (2003), this analysis found strong positive correlations between teacher resources, measured by the MKT survey, and the mathematical quality of their instruction. Table 4 shows Spearman rank order correlations that provide evidence on this matter. An examination of the first two columns shows that survey and lesson-based estimates of teachers' mathematical knowledge are correlated at .72, the lesson-based estimate of mathematical knowledge and MQI are correlated at .90, and lesson MQI is correlated with the survey measure at .58. All are significant ($p < .01$).

Table 3
Value-Added Ranks for Teachers at Participating Schools

				Value-Added Descriptives			Value-Added Percentile Ranks		
Teacher	MKT Percentile	MQI	Lesson MKT	*n*	FRL (%)	Raw Gain	Model 1	Model 2	Model 3
Tammy	35	1.67	2.00	67	100	42	97	97	100
Beryl	1	1.80	1.80	43	100	22	36	19	38
Cristobal	33	1.75	2.00	79	100	19	86	75	96
Paloma	7	1.33	1.33	64	100	23	48	26	65
Dean	31	1.40	2.00	49	100	36	94	86	94
Irene	28	1.40	1.60	9	100	33	75	61	77
Marco	69	1.33	1.83	125	46	26	62	83	44
Florence	6	1.60	1.60	25	80	21	21	30	25
Alberto	49	1.33	1.50	92	65	13	31	51	25
Felix	48	1.83	2.00	126	64	16	33	53	30
Gordon	25	1.00	1.75	30	13	12	80	71	81
Fay	89	2.00	2.17	101	55	18	76	63	76
Vince	93	2.50	2.83	110	39	27	90	84	83
Josephine	42	2.17	2.50	112	59	29	82	71	86
Melissa	44	1.83	2.00	105	50	10	39	26	35
Dolly	94	2.25	2.50	116	41	36	95	93	95
Arthur	98	2.17	3.00	74	14	27	87	82	92
Gabrielle	50	2.17	2.17	95	52	22	89	87	87
Edouard	69	2.67	2.50	86	38	23	70	60	46
Helene	47	1.33	1.33	74	38	11	9	4	3
Andrea	66	2.17	2.17	58	24	26	93	92	82
Ingrid	7	2.00	1.83	70	56	10	40	33	46
Hanna	55	2.17	2.50	31	55	37	96	96	93
Chantal	12	1.00	1.00	35	37	20	70	61	72

Note. All teacher names are pseudonyms. MKT Percentile = participants' percentile ranking against the national sample; MQI = mathematical quality of instruction; Lesson MKT = mathematical knowledge for teaching (MKT) estimate based on observations; *n* = number of students used to calculate value-added ranks; FRL = percentage of students who qualified for free or reduced-price lunch; Raw Gain = the average gain on the standardized assessment from the previous year; Model 1 = simple model; Model 2 = model with school fixed effect; Model 3 = model adjusted for student background covariates.

Convergent and Discriminant Validity

To assess Inferences a and b of our first assumption, we compare observational, survey, and value-added scores for evidence of convergence, shown in Table 4. In general, the strongest correlate of value-added scores is the lesson-based estimate of MKT (r = .46–.66). As noted above, this measure differed from overall MQI in that observers were directed to estimate

Table 4

Correlations Among Teacher Knowledge, Instructional Measures, Student Outcomes, and Demographic Makeup of Teachers' Students

	Survey MKT	MQI	Lesson MKT	Simple Model	School Fixed Effects	Student Background Adjusted	Average Raw Gain
Survey MKT	1	.58**	.72**	.41*	.51*	.25	.27
MQI	.58**	1	.90**	.45*	.38	.36	.32
Lesson MKT	.72**	.90**	1	.66**	.61**	.58**	.46*
Students' average 06-07 math scores	.60**	.30	.48*	.53**	.50**	.36**	.27**
Proportion of gifted students	.45*	.14	.38	.49**	.44**	.34**	.10
Proportion of accelerated students	.65**	.43	.45*	.36**	.36**	.12	.14*
Proportion of FRL students	−.52**	−.21	−.27	−.25**	−.13	−.10	−.04
Proportion of ELL students	−.60**	−.36	−.36	−.30**	−.17*	−.13	−.04
Proportion of SPED students	−.42*	−.07	−.17	−.35**	−.36**	−.32**	−.06

Note. Spearman rank order correlations are reported. All observational measures $n = 24$; all others $N = 222$. MKT = participants' normalized mathematical knowledge for teaching (MKT) ranking against the national sample; MQI = mathematical quality of instruction; Lesson MKT = MKT estimate based on observations. Accelerated students include those enrolled in algebra or enriched classes.
*$p < .05$. **$p < .01$.

the level of mathematical knowledge held by the teacher. Survey-based MKT and lesson MQI had lower correlations with student outcomes, in the range of .25 to .51.

Some may take these correlations to be one piece of evidence for the validity of value-added scores. However, critics of accountability systems based on value-added models might argue that these relationships result at least in part from the matching of better-qualified teachers with more able students. Comparing value-added models that do not control for student background and school to those that do yields insight into this claim. Results from the simple model and the school fixed effects model show roughly the same relationship to the instruction and knowledge variables. However, controlling for student-level covariates in the student background adjusted model reduces the correlation between the external predictors and outcomes.

To continue this line of inquiry, we investigate Inference c of our first assumption, that teachers' value-added scores should show little relationship to student characteristics. Table 4 shows that correlations between value-added scores and student background characteristics vary by model but are generally only modestly lower than the correlations among MQI, MKT, and value-added scores. Despite controls for student 2006–2007 SMA scores in each value-added model, student 2006–2007 average scores remain strong positive predictors of teachers' current-year value-added scores (r = .27–.53, $p < .0001$). Put plainly, teachers with more able students, measured by average SMA scores at entry into their classroom, have, on average, higher value-added scores. Teacher ranks in the simple model also correlate moderately with several other student background variables—positively with proportion of accelerated or gifted students and negatively with proportion of students who are eligible for FRL, ELL students, and SPED students. These relationships were weaker but still largely extant in the school fixed effects model, which corrects for between-school sorting of students and teachers but not for within-school sorting. This suggests either within-school sorting of more able teachers to more able students or bias in estimating teacher value-added scores. In the student background adjusted model, FRL, ELL, and accelerated status became uncorrelated with teacher scores, but the other student characteristics remained significant—despite the fact that the model itself controls for these descriptors at the student level.

There are two potential reasons for the lack of discriminant validity we find with scores from these three value-added models. The first is that these covariate-adjusted value-added models may inadequately control for student characteristics, either because the measures of student characteristics are crude (e.g., FRL status) or because student-level variability can best be accounted for in models with student fixed effects (McCaffrey et al., 2009). The second is, again, that higher quality teachers tend to be matched with high-achieving students as seen in Table 4. In both cases, it is possible that two teachers with similar-quality instruction and knowledge have different value-added scores owing to the students who populate their classes.

To continue to investigate this issue, we calculated partial correlations between value-added scores and teachers' average student characteristics after adjusting for MKT and MQI. These partial correlations are given in Table 5. These partial correlations are not statistically significant and do not provide evidence of an association between teacher value-added scores and the makeup of teachers' students after adjusting for teacher instructional quality and knowledge. However, they provide no evidence against discriminant validity; the lack of statistical significance may be due to insufficient statistical power, and in fact a number of these correlations are of moderate size.

To gain insight into the effects of instructional quality and a teacher's student population on value-added scores, we also calculated partial

Table 5

Spearman Partial Correlations Between Teacher-Level Compositional Variables and Value-Added Scores After Adjusting for Teacher Observational Measures

	Simple Model	School Fixed Effects	Student Background Adjusted	Average Raw Gain
Students' average prior-year math scores	.28	.23	.12	.11
Proportion of gifted students	−.06	−.11	−.04	−.41
Proportion of accelerated students	.40	.41	.14	.33
Proportion of FRL students	.07	.09	.17	.39
Proportion of ELL students	−.02	−.03	.07	.18
Proportion of SPED students	−.23	−.11	−.24	.05

Note. FRL = free or reduced-price lunch; ELL = English language learner; SPED = special education. Teacher observational measures include mathematical knowledge for teaching (MKT), lesson MKT, and mathematical quality of instruction.

correlations between observational and value-added scores controlling for the characteristics of each teacher's students.[5] The student composition variables controlled for were the average of a given teacher's current year's students on the prior-year SMA, percentage FRL, percentage ELL, percentage SPED, percentage gifted, and percentage of students in accelerated mathematics, algebra, or enriched mathematics. Results are shown in Table 6. Partial correlations using scores derived from school fixed effects model are not statistically significant, likely due to a lack of statistical power and the collinearity among teacher quality, school assignment, and student characteristics. However, after adjusting for teacher-level compositional variables, there is a significant correlation among teachers' lesson MKT, MQI, and the scores from the simple and the student background adjusted models. This implies that the relationship between lesson MKT and MQI and value-added scores for the two models without school fixed effects is not simply due to the differences in the makeup of teachers' students; there is a teacher quality "signal" in the scores even after controlling for the students assigned to specific teachers.

Consequential Validity

Next we investigate our second assumption, that decisions based on value-added scores will not create negative consequences for those working within a school or system. Our first inference states that value-added scores can be used to identify both excellent and poor teachers accurately. Defining excellent and poor is, of course, subjective. We rely on our discipline-grounded observational rubric but recognize that others may have different views. Determining how accurately these groups should be identified is also

Table 6
Spearman Partial Correlations Between Teacher Observational Measures and Value-Added Scores

	Simple Model	School Fixed Effects	Student Background Adjusted	Average Raw Gain
MKT	.16	.22	.19	.57*
MQI	.56*	.30	.52*	.42
Lesson MKT	.57*	.39	.55*	.46

Note. MKT = mathematical knowledge for teaching; MQI = mathematical quality of instruction; Lesson MKT = lessson-based guess at teacher's MKT. Partial correlations reported after adjusting for the average of the current years' student 2006–2007 state mathematics assessment, percentage of students who qualified for free or reduced-price lunch, percentage of students who were English language learners, percentage of students in special education, percentage of students designated as gifted, and percentage of students in algebra or accelerated or enriched mathematics.
*$p < .05$.

problematic, for there is no agreed-on metric or threshold, or even any academic discussion of either for the level of accuracy needed for accountability schemes. To demonstrate how this might be done and to explore this last interpretation of scores, we take various approaches.

Table 7 illustrates one approach. It shows the percentage of teacher-level variation explained by each predictor, entered alone, into a model with a teacher, school, and classroom random effect. If observational or survey measures were tightly aligned with teachers' value-added scores, these predictors would explain a significant proportion of the teacher-level variance. However, Table 7 shows that lesson-based MKT explains 46% of variation in scores and survey-based MKT explains roughly 37%; lesson MQI accounts for only 11%. Prior-year value-added scores, to date the best predictor of teachers' future performance (Gordon et al., 2006), account for between 32% and 45% of variation in scores. Most variation in teacher scores is unexplained in these models.

A second approach involves plotting the data to examine the extent to which excellent and poor teachers are similarly identified by observational or survey and value-added methods. Figure 1 shows teachers' value-added score, estimated in percentile units in the simple model, plotted against MQI. This figure demonstrates several points. First, there are no teachers who have an above-average MQI but low value-added score, easing concerns about unfair dismissals of low-value-added but high-MQI teachers. Second, all high-MQI teachers also have high value-added scores, suggesting that these teachers would be accurately rewarded. However, five teachers have value-added rankings above the 60th percentile—high enough to be rewarded in several districts, including Houston—but have MQI scores

Table 7
Reduction in Teacher-Level Variance Due to Teacher-Level Variables

Effect	Estimated Slope	Standard Error	Degrees of Freedom	t	Variance Reduction (%)
MKT	2.87	1.10	1667	2.60	37**
MQI	4.62	2.58	1667	1.79	11
Lesson MKT	7.57	2.08	1667	3.64	46**
Simple model 06-07 value-added score	4.02	1.39	1504	2.88	33**
School fixed effect model 06-07 value-added score	3.95	1.48	1504	2.67	45**
Student background adjusted model 06-07 value-added score	3.64	1.22	1503	2.98	32**

Note. t = estimated slope/standard error; MKT = mathematical knowledge for teaching; MQI = mathematical quality of instruction; Lesson MKT = lessson-based guess at teacher's MKT. A hierarchical linear model (HLM) was originally fit on the full districtwide data with random effects for school, teacher, and classroom and adjusting for the students' previous-year state mathematics assessment (a third degree polynomial), standardized reading scale, and standardized science scale scores and grade. The marginal residuals from that model were saved. A baseline HLM model with school, teacher, and classroom random effects were fit to the subset of data taught by teachers in our sample using the marginal residuals as an outcome variable. Then separate HLMs were fit adding each of the effects above to the baseline model. The estimates in the table are parameter estimates for those teacher level effects in the HLMs. The variance reduction is the percentage drop in teacher level variance when the effect is added to the baseline model. This two-stage process was utilized so that the parameters for adjusting for the student covariates could be estimated using the districtwide data. This yields a more accurate estimation of the proportion of variance accounted for statistics than if only the reduced data were utilized in the estimation.
**$p < .01$.

below 1.5; this means that half or more of their lessons were judged to be of low quality. These teachers compose roughly one fifth of our sample. Thus, while all high-MQI teachers would be rewarded, several low-MQI teachers would be similarly rewarded.

We also ask how accurately value-added scores identify poor teaching. Looking again at Figure 1, we see that eight teachers have an overall MQI of less than 1.5, meaning that observers rated more than half of their lessons as having low MQI. This means that their instruction was determined to be significantly problematic and possibly requiring intervention, with very high rates of mathematical errors and/or disorganized presentations of mathematical content. Yet only one of those teachers, Helene, is identified as failing in the value-added model. While recommending one poorly performing teacher for remediation or removal would be a net benefit over the current system, which identifies very few such teachers (Weisberg, Sexton, Mulhern,

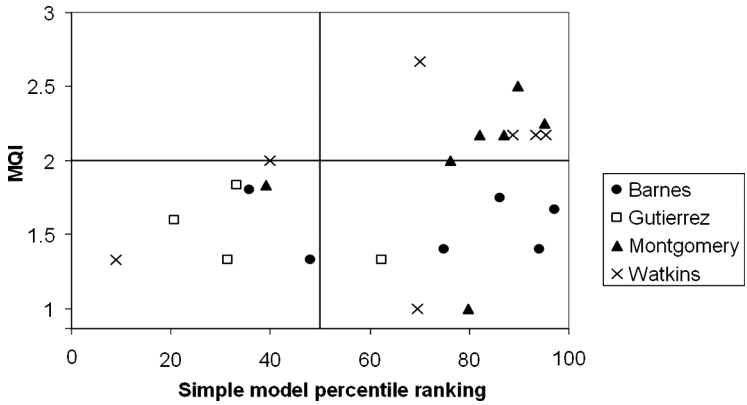

Figure 1. **Mathematical quality of instruction (MQI) versus value-added score from the simple model.**

& Keeling, 2009), the failure to identify seven others suggests that many students will continue to study mathematics in less than optimal environments.

There are several reasons that teachers with low MQI scores might have relatively high value-added scores. Interviews with teachers and school staff revealed that after-school tutoring and/or mathematics enrichment programs occurred in each school; there is also the possibility that parents contribute unevenly within and across schools and years to student mathematical learning (see Ishii & Rivkin, 2009). Theoretically, tutoring and other educational inputs could be included in value-added modeling, but in this district, and we suspect others, information on these inputs is not routinely collected on an individual basis. Furthermore, test preparation and discussion of the test were occasional topics in the videotaped lessons, but unevenly so among teachers. Teachers who engage in extensive test preparation are thought to post higher scores even in the absence of significantly better student understanding (Koretz, 2008).

Finally, it is also possible that this mismatch between value-added scores and MQI might be attributed to the mathematical emphasis of the MQI and a corresponding lack of focus on other facets of classrooms (e.g., climate). To address this possibility, we reexamined the five cases of teachers with value-added scores greater than the 60th percentile but with MQI scores less than 1.5 (see Figure 1) with the intent of developing case studies. From these five cases, we chose two with the lowest MQI scores to explore, asking whether there are facets of instruction not captured by the MQI that might have produced strong student achievement. Thus, the case studies serve as a check on our results. These case studies also help address two

other key issues. First, examining specific teachers' instruction in detail provides the reader with a gauge of how problematic the instruction of this subset of our sample was. Accidentally rewarding teachers who were simply mediocre in a sample of outstanding educators would be different, from a policy perspective, than accidentally rewarding teachers whose instruction appeared to be harmful to students. Second, examining teachers in detail provokes a discussion about the implications for accidental reward of teachers under value-added reward systems. Because the high-value-added and low-MQI group constitutes a substantial pool of teachers, it seems likely that such errors would be detected by colleagues and superiors within schools.

We reexamined each teacher's lesson, reviewed his or her performance on the MKT assessment, and read through postlesson and general interviews. We discussed each case in light of this intensive examination of the data, drafted memos, and tested our assertions against other raters' perceptions. Both teachers work in the two middle socioeconomic status schools in our study and use the same set of curriculum materials, *Connected Mathematics* (CMP), facilitating the comparison.

Case Studies

Chantal. Chantal is our first case study.[6] On paper, she does not appear to be particularly well prepared to teach middle school mathematics. Although she has 8 years of teaching experience, she is a generalist, teaching both another subject as well as general seventh grade mathematics. She holds an elementary certification and taught elementary school in the past. Like many elementary-certified teachers, she does not have a degree in mathematics and took only a handful of mathematics classes while in college. Her paper-and-pencil MKT assessment put her in the 12th percentile of our national sample, and observations of her classroom put her at the very bottom for MQI (1.00) and lesson-based MKT (1.00).

Chantal's instruction does have strengths. In an interview she reports that she often uses questions such as "Who agrees with that? . . . Who disagrees? Why do you disagree?" and our observations confirmed this as a frequent instructional strategy. During one lesson, she orchestrates a discussion during which students present different methods for solving the same problem. She seems to care about her students and is interested in both mathematics and the work of teaching. She also holds equitable views of mathematics learning, volunteering during an interview that "I think that all kids are able to do math, so for me I think that they're just different characters. Just different interesting situations." She speaks well of recent professional development opportunities and generally has a positive attitude about her work.

However, there were many times where Chantal's instruction clouded rather than clarified the mathematics of the lesson. A segment of a lesson taken from the "Comparing and Scaling" unit in CMP illustrates this pattern. She begins by reading the problem aloud:

> Dario has two options for buying boxes of pasta. At Corner Market he can buy seven boxes of pasta for six dollars. . . . At Super Foods he can buy six boxes of pasta for five dollars. . . . At Corner Market he divided seven by six and got [1.166667]. He then divided six by seven and got [0.85714286]. He was confused. What do these numbers tell him about the price of boxes of pasta at Corner Market?

The purpose of this introductory problem is to help students think about the two unit rates associated with the cost of pasta at Corner Market (7 boxes for \$6) by interpreting Dario's two division problems within the context given. In this case $7 \div 6 \approx 1.17$ corresponds to the number of boxes one can purchase for a single dollar and $6 \div 7 \approx 0.86$ corresponds to the number of dollars required for a single box.

Chantal rereads the question, "What do these numbers tell him about the price of boxes of pasta at Corner Market?" and Shawn responds, "They're expensive?" Shawn's answer suggests that he does not recognize that each division represents a different unit rate associated with the same store's pricing scheme. Chantal's answer does not move productively toward resolution:

> But he got two different prices, Shawn, so why would you say they're expensive with two different prices? One seems to be about eighty-six cents [\$0.86] and one is a dollar seventeen [\$1.17]. So are they expensive or do we really know?

In fact, Chantal incorrectly interprets the two numbers as two different prices per box. Not surprisingly, during the student work time that follows, most believe that both \$0.86 and \$1.17 represent prices per box, and many continue to try to explain why pasta is so expensive. At this point, the lesson has moved off track; students are not making mathematical progress, and Chantal's interventions do little to correct the situation. This is not a surprise: She has given students no tools to think about the units involved, namely, dollars per box or boxes per dollar, and has in fact misled them about the correct units. Drawing the class together, Chantal steers students toward her own understanding of the situation:

> Chantal: So what does this answer tell me, [1.17]. Is it the price or the box?
> Students: The price.
> Chantal: It's the price?
> Students: It's both.
> Chantal: OK, how many of you agree it's price? How many of you think its boxes?
> Student: I think it's boxes.

> Chantal: How many of you agree it's the price for a box? How many of you agree you don't know what it is? [many students raise their hands]. That's a good answer. At least we're on one page.

This passage reveals a key problem that plagues all of Chantal's lessons on rates. When she asks whether 1.17 is the "price" or "box," she means to ask whether 1.17 corresponds to "the price (i.e., number of dollars) for one box" or "the number of boxes for one dollar" (respectively). However, here as in many other places in this and similar lessons, she leaves the "for one unit" implicit in these unit rates, a move that might confuse students just beginning to learn about this topic. She continues,

> Chantal: What it's showing me—this [pointing to 1.17] is actually price. So each box is going to cost what?
> Student: A dollar seventeen [$1.17].
> Chantal: Yay, somebody knew . . . a dollar seventeen is going to be the price. This [pointing to 6/7] is going to tell me box for price, is that [6/7] really how we usually figure that out?
> Student: Yes.
> Chantal: If we want to know the price of a box, one box, how do we usually do that?
> Student: We divide the boxes by the price.
> Chantal: We divide the boxes by the price, right, to find a unit rate. Because we want to know how much one box costs, so am I going to divide the box by the price, the price by the box?
> Student: No.
> Chantal: So, I'm going to use this one, correct [pointing to 7/6 = 1.17]? And that's going to tell me how many one costs? Okay. So this is kind of like he shouldn't have used this [pointing to 6/7 = 0.86 then crossing it off].

This short passage is filled with teacher errors. She labels 1.17 as the price when in fact it is the number of boxes per dollar. She identifies 0.86 as the "box for price," perhaps meaning number of boxes for each dollar, but 0.86 is in fact the price per box. Chantal also agrees with a student who suggests an incorrect procedure for finding the price of a box. Finally, she crosses out the computation that led to 0.86, suggesting that it is nonsensical when it is in fact the correct answer to the question she had just posed.

The lesson continues in much the same vein. When students struggle to make sense of her mathematics, she chides them for failing to listen. Later, she suggests that in looking for a unit price "it's usually a whole divided by another whole of a price." Remarkably, some students appear to recover and make sense of the problems in the curriculum materials.

This vignette is representative of the set of lessons we observed Chantal teach. In every one, there were significant problems with the basic mathematics of middle school. She reasons incorrectly about unit rates. She

concludes that an answer of 0.28 minutes must actually be 0.28 seconds because one cannot have a fraction of a minute. She tells students that integers include fractions. She reads a problem out of the text as 3/8 + 2/7 but then writes it on the board and solves it as 3.8 + 2.7. She calls the commutative property the community property. She says proportion when she means ratio. She talks about denominators being equivalent when she means the fractions are equivalent.

These features of her instruction were independently noted by all observers and resulted in the lowest possible MQI and lesson-based MKT estimates. In contrast to other teachers, few of whom struggled to answer the *student* problems posed by the materials, Chantal made at least one major computational or conceptual error per lesson when presenting student-level work. There was also little evidence of specialized knowledge for teaching that would help her convey complex mathematical ideas in ways that would be usable for learners. And in every lesson, student behavior is a significant issue. Whether students sense her lack of command over the mathematics and respond accordingly or simply react negatively to Chantal's tough-love style of behavior management, teacher and students engage in constant direct confrontation.

We examined this case to determine whether Chantal's instruction may enhance student learning in ways not captured by the overall MQI score. As noted above, Chantal's instruction does have strengths, notably the ways she asks students to participate in the construction of mathematical knowledge. A subscale on the more detailed MQI instrument measured the level of student participation in the construction of mathematics; its relationship to value-added scores was modest (.04–.20, $p > .10$), suggesting this was not a strong factor in student outcomes. She also begins each class period with a 5-minute computational review and practice session, which may well have contributed to her students' strong performance on the state assessment. However, we observed other teachers with higher than expected value-added scores who did not use this teaching strategy and teachers who used this strategy who did not have high value-added scores. Although we cannot rule this out as a causal agent, it does not seem likely that computational practice is the sole cause of her strong value-added scores. Other than the student practice, we could find no other evidence in the six recorded lessons of features of instruction that could reasonably be hypothesized to contribute to student achievement.

Gordon. Gordon, our second case study, is in his 4th year of teaching. Since beginning his teaching career, Gordon has worked in three separate schools, including one high school. He had a long career in a related field prior to returning to school and receiving a mathematics-specific teaching certification. While he does not hold a mathematics degree, he describes extensive coursework in mathematics; his experience and training indicate

potential for substantial mathematical knowledge. However, Gordon scores in the 25th percentile of the national MKT sample, and in an interview he noted that he has difficulty predicting student errors. He also appeared puzzled by MKT items that asked him to work with student thinking and nonstandard solution methods. Gordon was tied for last (with Chantal) on MQI (1.00) and was in the bottom third for lesson-based MKT (1.75).

A review of the six captured lessons suggested there are positive elements to Gordon's classroom. While he occasionally makes serious mathematical errors and frequently makes minor errors, most are computational and pedagogical rather than conceptual, and he and his students quickly correct them. Students, many of whom are classified as accelerated learners, show evidence of serious mathematical thinking in that they are quick to ask questions, make mathematical observations, or even request more difficult problems. Gordon circulates among students, attempting to keep students on task, checking work, and answering questions.

However, the overwhelming impression of Gordon's classroom is that there is very little mathematics occurring. In many lessons, Gordon offers only the briefest of mathematical presentations, typically referring students to the text and assigning a series of problems. In one lesson, he fails altogether to directly teach any material. And throughout the classes we observed, student behavior is a serious issue. A constant stream of chatter—much of it off topic and occasionally stoked by Gordon—permeates student work time, which in turn constitutes most of each class.

These patterns are seen in the first lesson we observed. The lesson begins with roughly 20 minutes of homework review. Students call out answers to the problems assigned from the textbook, and Gordon verifies or corrects these answers. Then amid student chatter, he hands out a bag containing colored chips to a trio of students, saying,

> Okay, pick it out, pick out about five or six and then try to guess the colors that are in there, how many of each. Pick one, and then put it back in the bag after you pick it.

Even in the first moment of instruction, we can see the task Gordon offers is not well posed. In fact, there are several potential mathematical ideas that might be explored through this task: Students might use the experimental probabilities to estimate the relative proportions of each color chip in the bag, they might compare empirical and theoretical probability, or they might notice that the empirical probabilities are more likely to be close to their corresponding theoretical probabilities as the number of trials increases. As it stands, however, the directions he provides give no hint as to the mathematical goal of the task, nor can students even answer his actual question without information about the number of chips in the bag. Furthermore, few students are even paying attention to the description of the task. As

classroom chatter rises, he restates the launch just a moment later for a different group of students:

> Student: What do we do?
> Gordon: You pick one—shh!—one of you has to record, the other one pick one out, and guess, after you pick about five or six, try to guess what the colors are inside, how many. But don't look!

This second iteration of his directions omit the instruction to pick one, record the color, and then replace it in the bag before drawing and recording another. This leads to the majority of students choosing five consecutive chips from the bag without replacement. Gordon responds first to a small group, then to the class:

> Gordon: You messed it up, too. You guys don't listen. One at a time. And then you—
> Student: I did take them out one at a time.
> Gordon: But you didn't put them back.
> Student: Oh, we're supposed to put them back?
> Gordon: You put them back.
> Student: You didn't say that.
> Gordon: Yes, I did. You don't listen.
> Student: You never said to put them back.
> Gordon: Yes, put them back! [Addressing whole class] Okay quiet. . . . What do you think happens to the experiment if you put 'em—if you take 4 or 5 out and you don't put them back and you pick one?
> Student: Aum, you cheat.
> Gordon: You've ruined the experiment.
> Student: You're supposed to pick one?
> Gordon: Yes, you take one, only one.

While failing to replace chips changes the task from the one he intended, the experiment they actually perform could be a legitimate one (i.e., probabilities without replacement). Without having a mathematical reason for the experiment he asks them to perform, however, there is no way for the students to understand why this variation matters. In describing the experiment as "ruined," he neither explains nor presses students to explain how the probability of selecting a given color changes with replacement or without replacement.

Initially Gordon suggests that students conduct "five or six" trials, but later he suggests that they continue drawing until they have reached 30 trials. He says, "Try to take a guess what you have. And if you're not sure, keep picking one out at a time." Yet he never discusses the key idea that the likelihood that the experimental results resemble the theoretical probabilities increases with the number of trials. None of the potential mathematics that can be drawn out of this experiment is ever revealed, and in fact the decision about the number of trials is made to seem arbitrary.

As the class continues, Gordon asks students to look in their bags and then determine the experimental and theoretical probabilities for choosing each color. When students struggle to calculate the theoretical probability, he defines it, saying, "That's real. Theoretical is real. What you should get," and to another student, "How many you had of each [color in the bag]." Although evidence from the lesson suggests that Gordon clearly understands the difference between theoretical and experimental probabilities, his use of mathematical language is problematic. "Real" is not a good synonym for "theoretical probability," as "real" might as easily be applied to what was drawn (what students really got) as to what was in the bag. "What you should get" does not provide students a usable definition for this term. "How many of each" is simply incorrect, for it fails to take into account the relationship between the number of each color chips and total number of chips. Yet here, as in other lessons, students appear to master the material despite his poor articulation of key concepts.

Other lessons follow a similar pattern. Class begins with a recitation of answers to the homework and an occasional perfunctory solving of a troublesome problem. He then launches a task, often by assigning pages from the book or by providing a brief, unclear set of directions. Student chatter is constant, and Gordon spends a large portion of class time repairing the initial launch of the task and managing behavior issues. There is rarely any summarization or closure to his lessons and very little active teaching. His interactions with students are also notable for his frequent inability to follow student thinking. Often he elicits student methods but provides no feedback, as if he had not heard them. In other cases he provides confusing feedback or mistakenly evaluates student work.

We again searched for evidence that factors other than the MQI might have caused high value-added scores. Gordon's instruction did contain several at least neutral elements, including the fact that most errors tended to be computational rather than conceptual and that he provided routine supervision of and feedback on student work. However, these were not features unique to Gordon's teaching, suggesting that they were not drivers of his high value-added score. Instead, we conjecture that it might be the student population in his class—accelerated, highly motivated—that produced the scores shown in Table 3. Nevertheless, given this research design we cannot definitively identify the reason his value-added scores are so high.

Case Discussion

Based on these analyses, we might expect Chantal's and Gordon's value-added scores to be in the bottom quintile. However, Chantal is placed in the high second quintile according to two models, and Gordon's performance is in the top quintile. Equally troubling is the fact that these two teachers' performances are indistinguishable, in terms of value-added scores, from

teachers whom expert raters viewed as highly accomplished, including Vince and Arthur (see Table 3).[7] A review of the three other teachers in this high-value-added and low-MQI category suggests that their lessons were, in many ways, similar to Gordon's. While none featured major conceptual errors on the part of the teacher, two (Dean and Irene) featured frequent mathematical imprecision, one teacher provided little actual instruction (Marco), and in all three cases the pace of instruction was slow.

As suggested when we introduced the case studies, one can use these cases to gauge the seriousness of the error that would arise if Chantal and Gordon were rewarded based on their value-added scores. If these teachers' instruction was simply mediocre, we conjecture that policymakers (and parents) might not object to their inclusion in the pool of rewarded teachers. However, in both cases the instruction appeared potentially harmful to students' learning, and rewarding such teaching would very likely be perceived as inappropriate by policymakers as well as other teachers and school administrators.

While we could not locate studies of potential consequences stemming from the inaccurate reward of teachers, there is reason to think that these consequences may be serious. To the extent that both teaching quality and value-added-based rewards are public within schools, distorted incentives might result. For teachers motivated by pay-for-performance incentives, observing Gordon might lead to competition for teaching accelerated classes. The trend in the larger district data set toward special education teachers having lower value-added scores, even correcting for student composition, also suggests that teachers may avoid serving this important population. Value-added rewards may affect teacher morale and cooperation as well, decreasing the incentives for teachers to collaborate to solve individual student learning difficulties, plan instruction, and improve practice.

These case studies also help address concerns about the adequacy of our observational instrument for detecting teacher quality. Our MQI coding suggested that the mathematics presented by teachers like Chantal and Gordon is not likely to be the agent causing strong student gain. Our in-depth case study of their instruction and our more cursory review of cases like Dean, Irene, and Marco suggest that there were no other obvious teaching characteristics that could explain the strong gains. Instead, it seems possible either forces external to these teachers—their teaching assignments or the compensatory parental inputs as suggested in Ishii and Rivkin (2009)—or stochastic variation is responsible for their scores. Furthermore, even if teachers did contribute causally to their students' strong scores, our evidence suggests that the pathway would be through avenues unrelated to the basic quality of instruction—extensive test preparation activities, for instance.

Our quantitative results, bolstered by the detailed information we gain from the case studies, suggest that value-added scores alone may not be accurate indicators of high- and low-quality teaching for accountability

purposes. While all teachers judged to have above-average instruction also had above-average value-added scores, five teachers had low MQI scores but value-added scores greater than the 60th percentile, and only one of the eight very low MQI teachers, Helene, was consistently in the bottom quartile according to our value-added models. These observations lead us to conclude that value-added scores, at least in this district and using these not-uncommon models, are not sufficient to identify problematic and excellent teachers accurately, as stated in our second assumption, Inference a.

Conclusion

This study evaluated two propositions critical to the validity of current and proposed uses of value-added scores: that these scores converge with other indicators of instruction and diverge from theoretically unrelated constructs and that scores are sufficiently accurate to support decisions made in specific cases. We did find evidence that teachers' value-added scores from some models converged with expert ratings of their instruction, but we also uncovered evidence that these same scores correlated somewhat with aspects of the composition of students in a teacher's classroom. We also found that while a substantial number of teachers were classified similarly by their value-added and observational scores, a large minority were not.

These mixed results could have arisen for several reasons. Some might argue that our measure of high-quality instruction may be inadequate or too narrowly defined. However, we took care to deploy generally accepted criteria for evaluating teaching and to use case studies to check for strong teaching elements that our rubric might not have captured. Others might suggest that our lack of fit between observational measures and value-added scores can be explained by deficiencies in our statistical models. This is a potentially valid criticism; however, we argue that in practice districts across the country may have problems similar to those encountered in this study, including a lack of historical links between teachers and students. Given that many states have just enabled such links under Race to the Top, it stands to reason that historic data will not be available in many locations for several years. Furthermore, some districts have intentionally used covariate-adjusted models (e.g., Dallas). Finally, as models are refined and improved, it is incumbent on the research community to empirically investigate their validity for teacher accountability schemes, much as this study does with simpler models. These models may be better, statistically, but validity is not demonstrated by statistical superiority.

Given these results, how can value-added measures be used? We recognize that the current teacher evaluation system is insufficient for improving the quality of the workforce and in fact found evidence that value-added scores could potentially play a role in improving this system if used wisely. Value-added scores do show convergent and, to a lesser degree,

discriminant validity—that is, they do carry a "signal" about the quality of classroom instruction and thus may be valuable and relatively inexpensive tools. In this vein, we suggest districts use value-added scores in combination with high-quality, discriminating observational systems or as a trigger for such observational systems' use. In the latter scenario, for instance, high scores might prompt visits to about-to-be rewarded teachers' classrooms; low scores may flag teachers for additional data collection and possible remediation or termination. However, we doubt that using value-added scores in combination with existing teacher evaluation systems will yield accountability systems with adequate accuracy. Evidence that existing observational systems capture little variation in quality (Weisberg et al., 2009) suggests that value-added scores would drive observed variation in mixed-methods accountability systems. Better observational instruments are needed.

Although we do recommend the use of value-added scores in combination with discriminating observation systems, evidence presented here suggests that value-added scores alone are not sufficient to identify teachers for reward, remediation, or removal. Although our correlations are in the same range as those of other studies that have investigated the relationship between value-added scores and external criteria, there is still a significant amount of disagreement in the categorization of teachers as effective or not effective. This finding was further confirmed by the case studies, which suggested that high value-added teachers did not necessarily have strong instruction. These data also suggested that teachers who should be the target of remediation-type interventions would not be identified as such by value-added scores.

This study does not provide definitive answers to why some teachers had divergent instruction and student scores. To do so would entail an elaborate study measuring dozens of potential influences on student achievement; however, based on our analyses, several hypotheses for this divergence stand out. Gordon and other teachers with accelerated students outperformed their peers, even in models that controlled for this student characteristic. By contrast, many of the teachers ranked lowest in value-added scores were teaching classes classified as special education. This phenomenon might be partly explained by special education teachers' tendency to have lower MKT, as noted in Hill (2007), and the fact that families with special education students may have access to fewer resources to "compensate" for lower quality instruction as compared to families with accelerated or gifted students (Ishii & Rivkin, 2009). However, there remains concern that even models that control for student population still place these teachers in the lowest quartile. Whatever the explanation, this phenomenon poses a serious threat to accountability systems based on teacher value-added scores and, if generalizable to other districts, would create disincentives for teachers to teach the lowest-performing students.

Finally, this study highlights the importance of considering value-added scores from more than a statistical standpoint. The overwhelming majority of the debate regarding the use of value-added scores has been purely statistical; while valuable, such research cannot investigate the validity of teacher scores vis-à-vis planned uses of scores. We encourage more such research, as it is urgently needed.

Notes

The research reported in this article was funded by NSF Grant EHR-0335411 and WT Grant/Spencer Foundation Grant 200900175. We would like to thank an incredible district mathematics coordinator for her help in making this study possible and Dan Koretz and John Papay for a helpful read of an earlier draft. Eric Anderson provided valuable research assistance. Errors remain the property of the authors.

[1]Special education students were not included.

[2]Unfortunately, in 2005–2006, the data linking students to teachers were not complete for some of the schools. These schools were held out of the analysis in models that included teacher effects in that year.

[3]These teachers were included in this study. All were fluent English speakers, allowing us to collect interview and survey data without the use of a translator. Spanish-language lessons were coded by Spanish-speaking coders with the aid of a close translation of the transcript.

[4]This cutoff is arbitrary; however, it is close to the cutoff used in Houston, in which seven students per teacher are required before reports are generated.

[5]We thank an anonymous reviewer for suggesting this analysis.

[6]Chantal, like other teacher names in this article, is a pseudonym.

[7]A case study of Vince's teaching is available on request from the authors.

References

American Educational Research Association/American Psychological Association. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, *37*(2), 65–75.

Ball, D. L., & Forzani, F. M. (2007). What makes education research "educational"? *Educational Researcher*, *36*(9), 529–540.

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.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

Boyd, D., Lankford, R. H., Loeb, S., & Wycoff, J. (2003). Understanding teacher labor markets: Implications for educational equity. In M. L. Plecki & D. H. Monk (Eds.), *School finance and teacher quality: American Education Finance Association 2003 yearbook* (pp. 55–84). Larchmont, NY: Eye on Education.

Buddin, R., & Zamarro, G. (2008). *Teacher quality, teacher licensure tests and student achievement* (RAND working paper). Retrieved from http://www.rand.org/pubs/working_papers/2008/RAND_WR555.pdf

Center for Educator Compensation Reform. (2007). *The evolution of performance pay in Florida*. Retrieved from http://www.cecr.ed.gov/guides/summaries/FloridaCaseSummary.pdf

Center for Educator Compensation Reform. (2010). *Compensation reform in action*. Retrieved from http://cecr.ed.gov/reformInAction/

Clotfelter, C., Ladd, H. F., & Vigdor, J. (2004). *Teacher quality and minority achievement gaps*. Durham, NC: Terry Sanford Institute of Public Policy.

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

Colorado teacher-evaluation bill enacted. (2010, May 18). *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2010/05/19/32colorado.h29.html

Dillon, S. (2009, November 10). States compete for federal school dollars. *New York Times*. Retrieved from http://www.nytimes.com/2009/11/11/education/11educ.html?_r=1&sq=colorado%20education&st=cse&adxnnl=1&scp=5&adxnnlx=1277474473-7wHNM/8mH25t3oRTa27+vg

Dillon, S. (2010, May 31). States create flood of education bills. *New York Times*. Retrieved from http://www.nytimes.com/2010/06/01/education/01educ.html?scp=1&sq=colorado%20education&st=cse

Duncan, A. (2009). *The race to the top begins: Remarks by Secretary Arne Duncan*. Retrieved from http://www.ed.gov/news/speeches/2009/07/07242009.html

Editorial: The New Haven model. (2009, October 28). *New York Times*. Retrieved from http://www.nytimes.com/2010/05/03/opinion/03mon3.html?_r=1&scp=1&sq=Editorial:%20The%20New%20Haven%20Model&st=cse

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (Discussion paper 2006-01). Washington, DC: Brookings Institution.

Grubb, W. N. (2008). Multiple resources, multiple outcomes: Testing the "improved" school finance with NELS88. *American Educational Research Journal*, *45*(1), 104–144.

Hanushek, E. A. (2007). The single salary schedule and other issues of teacher pay. *Peabody Journal of Education*, *82*(4), 574–586.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality*. Unpublished manuscript.

Hill, H. C. (2007). Mathematical knowledge of middle school teachers: Implications for the No Child Left Behind Policy initiative. *Educational Evaluation and Policy Analysis*, *29*, 95–114.

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

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy*, *4*(4), 520–536.

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (Faculty Research Working Paper Series RWP05-040). Cambridge, MA: Harvard University, John F. Kennedy School of Government.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. (2004). Certification testing as an illustration of argument-based approach validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 135–170.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 18–64). Westport, CT: Praeger.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City*. Unpublished manuscript, Harvard University.

Kimball, S. M., White, B., & Milanowski, A. T. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54–78.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function* (Working Papers 708). Retrieved from http://econpapers.repec.org/paper/umcwpaper/0708.htm

Koedel, C., & Betts, J. R. (in press). Does student sorting invalidate value-added models of teacher-effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*.

Koretz, D. (2008). Limitations in the use of achievement tests. *Journal of Human Resources*, *27*(4), 752–777.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis*, *25*, 287–298.

Lewin, T. (2010). School chief dismisses 241 teachers in Washington. *New York Times*. Retrieved from www.nytimes.com/2010/07/24/education/24teachers.html

Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies*, *40*(6), 699–711.

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, *27*(3), 255–270.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. M. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, *32*(2), 125–150.

McCaffrey, D. F., Koretz, D., Lockwood, J. R., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

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.

Medley, D. M., & Coker, H. (1987). How valid are principals' judgments of teacher effectiveness? *Phi Delta Kappan*, *69*, 138–140.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33–53.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237–257.

Papay, J. P. (in press). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*.

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*(2), 365–397.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2010). *Information and employee evaluation: Evidence from a randomized intervention in public schools* (Working paper 16240). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w16240

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the "Prospects" study of elementary schools. *Teachers College Record*, *104*(8), 1525–1567.

Sanders, W. L. (2006, October). *Comparisons among various educational assessment value-added models*. Paper presented at the Power of Two—National Value-Added Conference, Columbus, OH.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville: University of Tennessee, Value-Added Research and Assessment Center.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system, a quantitative, outcomes-based approach to educational measurement. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Sanders, W. L., & Wright, S. P. (2008). *A response to Amrein-Beardsley (2008) "Methodological concerns about the education value-added assessment system."* Retrieved from http://www.sas.com/govedu/edu/services/Sanders_Wright_response_to_Amrein-Beardsley_4_14_2008.pdf

Schacter, J., & Thum, Y. M. (2004). Paying for high and low-quality teaching. *Economics of Education Review*, *23*, 411–430.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., . . . Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11–35.

Weerasinghe, D. (2008). *How to compute school and classroom effective indices: The value-added model implemented in Dallas Independent School District*. Dallas, TX: Office of Institutional Research Dallas Independent School District. Retrieved July 16, 2010, from https://mydata.dallasisd.org/docs/CEI/SEI_CEI_Research.pdf

Weingarten, R. (2007, October 14). How to fix No Child Left Behind. *New York Times*, p. WK7.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: New Teacher Project.

Wiley, E. W., Spindler, E. R., & Subert, A. N. (2010). *Denver ProComp: An Outcomes Evaluation of Denver's Alternative Teacher Compensation System 2010 Report*. Boulder, CO: University of Boulder at Bolder, School of Education. Retrieved

from http://static.dpsk12.org/gems/newprocomp/ProCompOutcomesEvaluation April2010final.pdf

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *1*(1), 57–67.

Manuscript received November 13, 2009
Final revision received August 4, 2010
Accepted August 25, 2010