American Educational Research Journal
February 2014, Vol. 51, No. 1, pp. 73–112
DOI: 10.3102/0002831213517130

How Teacher Evaluation Methods Matter for Accountability: A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures

Douglas N. Harris
Tulane University
William K. Ingle
Bowling Green State University
Stacey A. Rutledge
Florida State University

Policymakers are revolutionizing teacher evaluation by attaching greater stakes to student test scores and observation-based teacher effectiveness measures, but relatively little is known about why they often differ so much. Quantitative analysis of thirty schools suggests that teacher value-added measures and informal principal evaluations are positively, but weakly, correlated. Qualitative analysis suggests that some principals give high value-added teachers low ratings because the teachers exert too little effort and are "lone wolves" who work in isolation and contribute little to the school community. The results suggest that the method of evaluation may not only affect which specific teachers are rewarded in the short

---

Douglas N. Harris is associate professor of economics and University Endowed Chair in Public Education at Tulane University, 302 Tilton Hall, New Orleans, LA 70118; e-mail: *dharri5@tulane.edu*. He also directs the Education Research Alliance for New Orleans that studies the effects of the city's post-Katrina school reforms. His research on education policy focuses on the teaching profession and market- and test-based accountability.

William K. Ingle is associate professor in the School of Educational Foundations, Leadership, and Policy at Bowling Green State University. His research focuses on human resource functions in education. His research interests also include the politics of school budget referenda and broad-based merit aid programs.

Stacey A. Rutledge is associate professor in educational leadership and educational policy studies at Florida State University. Her research explores how policies aimed at improving teaching and learning, such as test-based accountability and teacher quality, shape the work of district and school administrators, teachers, and ultimately, students' learning opportunities.

term, but shape the qualities of teacher and teaching students experience in the long term.

KEYWORDS: teacher quality, evaluation, accountability

Introduction

Policymakers are increasingly turning to evaluation and accountability for individual teachers as a way to improve school performance and student outcomes. The federal program, Race to the Top, requires participating states and school districts to measure and reward teachers and school leaders based on contributions to student achievement, or "value-added." Florida, for example, recently passed legislation requiring that teacher and school value-added comprise roughly 50% of the teacher evaluation, and these evaluations are the basis for high-stakes decisions about promotion, tenure, dismissal, and compensation. Many other Race to the Top states also allow schools to use locally developed measures of classroom and professional practice applied by either external peer evaluators and/or internal evaluators such as school principals. Informing this ongoing national experimentation, the Gates Foundation has invested $45 million in the Measures of Effective Teaching (MET) project that measures teacher effectiveness in many different ways, including student evaluations of teachers, student classroom work, and evaluations of classroom practice using multiple rubrics (e.g., Kane, McCaffrey, Miller, & Staiger, 2013). Behind these initiatives is a goal of not only improving teacher evaluation, but using this information to make high-stakes decisions about teachers' careers.

Little is understood about how such changes in teacher evaluation methods might influence not only the specific teachers rewarded in the short term, but the basic qualities and activities of the teacher workforce in the long term. If teacher evaluation is used to make hiring, promotion, tenure, and dismissal decisions—and if different evaluation tools give greater weight to some qualities over others—then the choice of evaluation tool would likely influence the qualities and activities of teachers. Indirectly, the shifts in incentives and evaluation measures could also influence who chooses to enter teaching as well as what types of preparation teachers can access. Our results suggest a less obvious implication: that teacher evaluation based on value-added is also likely to reduce emphasis on teachers' personal traits like sociability and ability to work well with multiple school actors—traits that school principals currently value highly, but that are more weakly related to teacher value-added.

Like value-added, the recent attention to teacher evaluations and their effectiveness is fairly new. While formal teacher evaluation tools have been in practice for decades, they give nearly all teachers the highest possible ratings and provide almost no information about the technical or

instructional core of teaching (Bidwell, 2001; Kennedy, 2004; Little, 2009; Parsons, 1960; Weisberg, Sexton, Mulhern, & Kelling, 2009). Principals' subjective conceptions of teacher characteristics (Ingle, Rutledge, & Bishop, 2011), low quality preparation (Elmore, 2000), views of teacher evaluation (Painter, 2000), and concerns for organizational cohesiveness (Marzano, Waters, & McNulty, 2011), as well as restrictive union rules that place the onus of the documentation of poor teaching practice on the principal, have all been identified as factors leading to weak teacher evaluations (Stodolsky, 1984). With studies increasingly concluding that some teachers are more successful in raising student achievement than others (e.g., Hanushek, 2011), more attention is being paid to the shortcomings of current evaluation practices and to the fact that evaluation results are largely ignored when making important personnel decisions about hiring, promotion, course assignment, termination, and compensation (Kennedy, 2010).

At the same time, there is growing agreement among policymakers and researchers alike that value-added measures by themselves are inadequate replacements for traditional teacher evaluation. This is partly why Race to the Top and related state policies require that student test scores be supplemented with other measures. Evaluations of teachers' classrooms by school principals and external peer reviewers are the most common additional metrics; however, as these multiple measures become more widely available, educators are finding that value-added measures often differ substantially from classroom observations and their own impressions of effectiveness (e.g., Jacob & Lefgren, 2008).

Existing data systems are, however, insufficient for understanding why the measures yield different conclusion about the effectiveness of individual teachers. Most districts still use formal evaluations that provide relatively little useful information about overall effectiveness and no information about the components of effectiveness that principals judge to be important. More extensive and detailed evaluations are being developed in many districts and states, but in those cases there are often legal impediments to obtaining formal, high-stakes evaluation scores. The Gates Foundation's Measures of Effective Teaching project is a partial exception, although even the extensive data collection for that project does not capture information from principals and was not designed to understand why any of the various metrics differ.

Principals' views are important because in the vast majority of schools they have long been responsible for conducting teacher evaluations (Liu & Johnson, 2006). In addition to both formal and informal observations of teachers in the classroom, principals receive feedback from students and parents and hear "water cooler" talk from other teachers. Charged with oversight of both teachers in their individual classrooms as well as the school as an organization, principals have a unique perspective on the contributions of teachers at their schools. On the other hand, formal evaluations by principals show less variability than almost anyone believes is credible (Weisberg

et al., 2009) and probably do not reflect principals' actual beliefs. Therefore, while it is clearly important to have a valid measure of what principals believe, we cannot rely on traditional formal evaluations and must try a different approach.

In this study, we draw on confidential principal interviews combined with value-added measures to address one main question: Why do teacher value-added measures differ from principals' impressions of teacher effectiveness? We answer this by comparing the teacher characteristics and skills associated with each effectiveness measure. After the literature review, we discuss how we collected our three linked sets of data from a midsized school district in Florida, Hillyer County (pseudonym). In addition to obtaining standardized tests annually in Grades 1 through 10 linked to teachers, interviewers asked each school principal to rate 10 teachers from their school on a prespecified range of characteristics, such as "strong teaching skills" and a "caring" personality, and to describe each of the 10 teachers in the principals' own words. The combination of closed- and open-ended questions provides a rich portrait of each teacher. More generally, the analysis highlights differing perspectives on the meaning of effectiveness and the characteristics associated with these diverse effectiveness measures.

In our mixed-methods analysis, we find some consistency in the teacher characteristics and skills associated with each effectiveness measure. However, there are also some noteworthy differences that provide a window into why they differ and, consequently, the types of teachers who would be rewarded under alternative accountability regimes. The open-ended responses provide additional depth to our understanding and highlight the importance of teachers' demonstrated effort and social interactions outside the classroom. As we show in the last section, these are critical issues informing the broader move toward teacher accountability policies and the choice between these and other types of evaluation techniques.

Theory and Literature Review

To understand how and why different measures of teacher effectiveness might vary, we begin with a general theoretical framework. Establishing a clear framework is complicated by the inconsistent use of terms like *effectiveness* and *performance* in the literature and the parlance of educators. *Effectiveness* is generally interpreted to mean influence on student outcomes, and in this respect, teacher value-added is a measure of effectiveness. The issue is less clear-cut with the principal ratings. We asked principals to rate teachers from "ineffective" to "exceptional," but their responses are likely to capture a combination of effectiveness, as typically defined, and their own notions of effectiveness. If principals want teachers to make contributions to the school community, then this might be considered unrelated to effectiveness, but even in that case it is reasonable to think that such

contributions outside the classroom have indirect influences on student outcomes (e.g., one teacher mentoring another could lead to better teaching and learning for the other teacher's students). For this reason, and to avoid overly cumbersome language later, we use *effectiveness* somewhat broadly and consider both our principal evaluations and value-added to be "effectiveness measures."

A similar problem arises with terms like *teacher quality*, which generally refers to teacher attributes (Fenstermacher & Richardson, 2007; Kennedy, 2008) that are thought to be associated with effectiveness. Other elements of teacher quality are personal resources (e.g., knowledge and credentials) and activities outside the classroom (e.g., collegiality and organization) (Kennedy, 2008). Moral traits such as honesty, compassion, and fairness might also be included (Fenstermacher & Richardson, 2007).

Our goal is not to argue that one of these ways of thinking about teacher quality is better than the others, but simply to clarify what this study is about. For this reason, we avoid the teacher quality terminology and instead refer to *predictors* of effectiveness or *characteristics* of teachers and teaching. Further, in trying to understand why the two effectiveness measures differ, we hypothesize the best predictors will differ across the two effectiveness measures—that the teacher characteristics associated with value-added are not the same as those associated with overall principal ratings. With this general terminology, we proceed by reviewing theory and evidence about differences between the two effectiveness measures and the roles of various predictors.

Theory and Evidence About the Relationship Between Different Evaluation Approaches

If teacher value-added and principal evaluation yielded exactly the same ratings of teachers, then there would be little point in considering how the characteristics of effective teachers might differ—the more closely related the effectiveness measures are, the more similar the characteristics associated with each effectiveness measure are likely to be. While prior research consistently shows that the two effectiveness measures are positively related, the correlations are weak enough that the characteristics distinguishing low- and high-rated teachers could differ. A number of these are older studies (Medley & Coker, 1987; Murnane, 1975; Peterson, 1987, 2000)[1] and are based on the relationship between teacher value-added and subjective teacher ratings that are from formal standards and extensive classroom observation (Gallagher, 2004; Kimball & Milanowksi, 2004; Milanowski, 2004). The most recent studies, most similar to our own, find correlations of .17 to .32 between teacher value-added and principals' informal evaluations of teachers (Jacob & Lefgren, 2008; Rockoff, Staiger, Kane, & Taylor, 2010).

Theoretically, there are many reasons why the two sets of effectiveness measures might differ this way. Of greatest interest here is that principals conceptualize teacher effectiveness as something other than simply raising student test scores, which may manifest itself through the characteristics of teachers whom they deem effective and ineffective. The level of stakes attached may also play a role. Campbell's Law states: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1976, p. 54). This means the measures might diverge because different stakes are attached, distorting one measure more than the other. A third factor is probably more important than the first two in statistical sense, but is also hardest to address (especially with these data): Measurement error in each measure reduces the maximum correlation to be much less than one.[2]

To isolate the distortion from the variation in effectiveness constructs, we focus on low-stakes effectiveness measures in our analysis to the extent possible. We also draw and build on prior research about the teacher characteristics associated with value-added and those valued by principals.

The Characteristics of High Value-Added Teachers

Earlier work in this area centered on rough signals of potential effectiveness like teacher education and experience, which are frequently available in administrative and national databases. With such a vast literature, it is useful to focus on numerous reviews, which have found mixed evidence regarding the relationship between teacher education and their contributions to student test scores (Harris & Rutledge, 2010; Rice, 2003; Wayne & Youngs, 2003; Wilson & Floden, 2003; Wilson, Floden, & Ferrini-Mundy, 2001), though there is some evidence that subject matter knowledge is important (Monk, 1994; Wilson & Floden, 2003). Teacher experience, in contrast, is the one factor that early evidence showed to be consistently and positively related to teacher performance (Harris & Sass, 2011; Rice, 2003).

Unfortunately, by failing to take into account *prior* student achievement, and only accounting for crude demographic measures, almost all the older studies incorporated in these reviews are essentially evaluating teachers based on end of year test scores, which mainly reflect what students bring to the classroom rather than what the current teachers and their attributes contribute to academic outcomes (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). By accounting for prior test scores, value-added measures help account for the nonrandom assignment of students to teachers and yield a less biased measure of teacher effectiveness (Guarino, Reckase, & Wooldridge, 2010; Harris & Sass, 2006; Kane & Staiger, 2008; Todd & Wolpin, 2003). More recent studies address this and other selection problems yet still come to similar conclusions (e.g., Aaronson, Barrow, & Sander, 2007;

Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004), although the returns to experience now seem to extend beyond the first few years (Harris & Sass, 2011) and some studies find evidence more supportive of credentials (Clotfelter, Ladd, & Vigdor, 2010). Content knowledge, pedagogy, and pedagogical content knowledge are surely important (e.g., Shulman, 1987), but these factors may not be instilled well in teachers through preparation.

Based on evidence from psychology (e.g., Schmidt & Hunter, 1998) as well as labor economics (Cunha, Heckman, & Schennach, 2010), personality seems to play a role in worker productivity. Borghans, ter Weel, and Weinberg (2008) theorize that different types of jobs require different combinations of personality traits, especially "directness" and "caring," and find evidence that some of these traits are correlated with productivity. This is perhaps not surprising, especially for jobs (e.g., teaching) that require substantial interpersonal interaction and communication.

Gallup's Teacher Perceiver Instrument (TPI) measures 12 themes drawn from research identifying the characteristics of teachers most successful at working with students. These themes include a candidate's capacity for the mission of student growth, empathy, rapport with students, individualized perception, listening, "investment" (satisfaction from the learner's response), "input drive" (capacity for seeking for new ideas and experiences to share with students), activation (capacity to motivate students), innovation (implementation of new ideas and techniques), "Gestalt" (a drive for perfectionism but works from individual to structure), objectivity, and focus (Metzger & Wu, 2008, p. 923). Metzger and Wu (2008) synthesize 24 studies from the psychology literature and conclude that the instrument gauges important teacher qualities through its affective themes (e.g., empathy) but only generally captures beliefs, attitudes, and values that principals desire (e.g., positive work ethic). More importantly, they find no link between the TPI and ratings from external evaluators (mostly trained educational researchers). Overall, prior research provides little evidence that specific characteristics are associated with either effectiveness measure.

Principals' Preferred Characteristics

Given our focus on explaining why teacher effectiveness measures differ, we are equally interested here in the relationships between teacher characteristics and principals' views of effectiveness. Principals play a critical personnel role in schools, and studies have found that they play an important, albeit indirect, role in improving student achievement at their school (Grissom & Loeb, 2011; Hallinger, 2005; Hallinger & Heck, 1996; Louis, Leithwood, Wahlstrom, & Anderson, 2010; Newmann, Smith, Allensworth, & Bryk, 2001). That role is only increasing and changing with the move to

more extensive, high-stakes classroom observations. Understanding the ways in which they choose, prioritize, and negotiate the characteristics of effective teachers has important implications on who gets hired and who stays in teaching.

In eight studies that asked principals to rank a prespecified list of teacher characteristics (Abernathy, Forsyth, & Mitchell, 2001; Braun, Willems, Brown, & Green, 1987; Broberg, 1987; Cain-Caston, 1999; Dunton, 2001; Harris, Rutledge, Ingle, & Thompson, 2010; Ralph, Kesten, Lang, & Smith, 1998; Theel & Tallerico, 2004), principals consistently report preferences for teachers who display strong communication skills (Braun et al., 1987; Broberg, 1987; Cain-Caston, 1999; Dunton, 2001; Ralph et al., 1998) and enthusiasm (Broberg, 1987; Dunton, 2001). Principals also report, although less consistently, preferences for teachers with certain teaching skills, teaching philosophies, types of knowledge, and an ability to work well with others. While only one of the studies considers the importance of whether teachers are "caring" (Harris et al., 2010), the authors find that this is the most important single characteristic, ahead of strong teaching skills and knowledge of subject matter. The same study finds that principals seek a "mixture" or a "balance" of personal and professional qualities when they select teachers.

As in the studies of student test scores, teacher credentials figure prominently when considering principal preferences. One group of studies focuses exclusively on the academic credentials of teachers who are hired. Using a nationally representative sample of recent college graduates, Ballou (1996) finds that applicants from more selective undergraduate institutions were no more likely to be hired than graduates of other institutions, a finding corroborated by Baker and Cooper (2005) in their analysis of a different national database (the Schools and Staffing Survey). Strauss and Vogt (2006) study the degree to which schools hire teachers who have strong academic credentials or who graduated high school in the same district. They find that schools located in communities with high levels of average adult education are more likely to hire teachers with stronger academic backgrounds and less likely to hire their own graduates.

Taken as a whole, this review suggests there are significant limits to what existing research can tell us about the characteristics of teachers associated with each effectiveness measure. The studies of student test scores, even when they have the data necessary to help account for selection bias, do not include the characteristics of teachers that principals think are important. Conversely, the characteristics principals say they prefer are almost never associated with any other measure of effectiveness. To understand why effectiveness measures differ, we therefore need a different approach to measuring teacher characteristics that might be associated with various effectiveness measures.

Data and Methods

Sample

We interviewed 30 principals from a midsized school district in Florida over a 2-year period during the summers of 2005 and 2006. The sample included principals of 18 elementary (or K–8 schools), 8 middle schools, and 4 high schools. Of the 30 schools represented in the sample, 10 were eligible for Federal Title I funds in the 2005–2006 academic year (8 elementary schools and 2 middle schools). The sample of principals is almost identical to the national average on race (sample district: 80% White; national: 82% White) and very similar in terms of the proportion with at least a master's degree or higher (sample district: 100%; national: 98.1%).[3] The sampled principals, however, are more likely to be female (sample district: 63%; national: 48%).[4]

While the sample of principals is diverse and reasonably similar in the nation's population of principals on the aforementioned important measures, this must still be viewed as a convenience sample. This choice is justified by the complex and sensitive nature of the data collection. We considered obtaining the formal evaluations of teachers that had been the basis for tenure and promotion decisions, but the district would not allow this, and more importantly, other evidence suggests that traditional formal teacher evaluations show unrealistically low variation in ratings (i.e., they are invalid measures of what principals consider to be effectiveness) (Weisberg et al., 2009).

We took several additional steps in obtaining principals' informal assessments of teachers, both to address the district's concerns about confidentiality and to address our own concern that principals might not be forthcoming about their actual views of individual teachers. In order to get open and honest responses from the principals, and to better understand their views, we therefore developed relationships with them over a 2-year period. Also, district personnel provided interview materials that allowed us to link informants' discussions of individual teachers to the district's administrative data that included test scores and teacher linkages—all the while maintaining teacher confidentiality. The choice of a convenience sample was therefore necessary to carry out the complex data collection.

Florida's accountability system gives grades to each school—from a high of A to a low of F— based primarily on student scores on math, reading, and writing on the state's standardized test, the Florida Comprehensive Assessments Test (FCAT). In addition to providing information to parents and voters, the grades are used as the basis for a formal structure of sanctions and rewards administered by the state government. While we do not explicitly consider other parts of the state's accountability system, it is noteworthy that Florida is considered to have one of the most aggressive systems in the

country (Carnoy & Loeb, 2003), and this continues to be the case with even more aggressive teacher accountability since our data collection was completed. In our earlier analyses of these principals we found that that school grades and the larger climate of accountability influenced principals' preferences for teachers (Rutledge, Harris, & Ingle, 2010). Given our small sample size, our analysis of the role of accountability in influencing principal responses is minimal, though there are some reasons to believe the accountability may have influenced interview responses.[5]

The Interviews

General Description

We conducted interviews with the principals over a 2-year period as part of a larger project about teachers. In the first interview, we asked principals about their practices and preferences in teacher screening and selection.[6] In the second interview, we asked principals to "rate each teacher on a scale from 1 to 9 with 1 being *not effective* to 9 being *exceptional*" and to describe the teachers in their schools in their own words as well as according to a prespecified list of characteristics we chose based on prior research. We piloted and improved both interview protocols with current and former principals external to the sample. The basis for this analysis is the second interview protocol, which is provided in its entirety as an appendix available from the authors upon request.

We began our interview with principals with several introductory questions. Then we gave principals a sealed envelope prepared by the district that in order to ensure confidentiality contained a list of 10 of their teachers with related identification numbers.[7] We then asked principals to complete three activities in which they rated the 10 teachers relative to each other on a scale of 1 (*low*) to 9 (*high*). In the first activity, we asked them to provide an overall effectiveness rating for each of the 10 teachers selected from their schools. Second, we asked them to rate each of the 10 teachers on the following selected personal and professional qualities: caring, communication skills, enthusiasm, intelligence, knowledge of subject, strong teaching skills, motivation, works well with grade team/department, works well with me (the principal), contributes to school activities beyond the classroom, and contributes to overall school community. The first seven characteristics in this list were found in the analysis of the first round of interviews to be among the most important characteristics that principals look for when hiring teachers (Harris et al., 2010). One characteristic from that study, "works well with others," was divided into two categories: works well with me and works well with team. After this rating activity, we asked them to explain why they gave these ratings to each teacher and to provide examples.

Through this design, we obtained not only numeric ratings for each of the teachers in the study, but also rich descriptions. In their open-ended responses, principals provided lengthy discussions of each teacher, explaining their ratings and providing specific examples of their general characteristics, strengths, and weaknesses. We draw heavily on these open-ended descriptions in our qualitative analysis when we compare the ways that principals described their highest and lowest ranked teachers to the ways that they described the high and low value-added teachers.

The interviews lasted an average of 1.5 to 2 hours. All interviews were recorded and transcribed. We coded and analyzed principals' responses using NVivo 6 and an iterative team memo-writing process (Miles & Huberman, 1994). We developed codes drawing from both the research on hiring and teacher effectiveness (e.g., prespecified teacher characteristics, such as "caring") as well as our own iterative and inductive process in which codes and themes emerged (e.g., "seeks professional development"). Principals' discussions of individual teachers and descriptors were coded drawing from our prespecified list of characteristics (e.g., caring, subject matter knowledge) and any other descriptors that principals mentioned. These discussions were coded as being positive, average/adequate, or negative.[8]

We placed teachers into "low" and "high" categories based on the effectiveness measure and then wrote memos based on principals' responses in different combinations. For conciseness, we refer throughout the remainder of the study to teachers who are high value-added (HVA), low value-added (LVA), high rating by the principal (HPR), and low rating by the principal (LPR). Memos were also written on individual characteristics (e.g., caring) in order to get a sense of how principals conceptualized these characteristics. Memos were written and revised several times until we had achieved theoretical and empirical saturation (Denzin & Lincoln, 1998).

We analyze the interview data in a variety of ways that are most relevant to our research question. Specifically, by comparing what principals say about teachers with different combinations of effectiveness measures, we can learn why the effectiveness measures themselves differ from one another, namely, the differences in the constructs being captured. Rather than rely solely on our prespecified list, we therefore quantified some of the qualitative data using the aforementioned coding system, allowing important teacher characteristics to emerge independent of our prespecified list.

Descriptive Statistics for Quantitative Data From Interviews

The descriptive statistics of the overall and prespecified characteristics ratings are shown in Table 1. The mean and standard deviation of the overall teacher ratings by principals partially mask the skewed distribution of ratings. Table 2 shows the distribution of ratings by principals of teachers'

overall effectiveness by rating and school level. Sixty-nine percent of the teachers are rated as being in the top three categories, while 26% and 4% are in the middle and bottom third, respectively. High school principals tended to rate their teachers lower than the elementary and middle school principals.

The uneven distribution was expected, given past evidence that principals tend to give quite high ratings to large percentages of their teachers.[9] The same skewed distribution arises with the characteristic measures and we therefore report only nonparametric chi-square tests of statistical significance.

Student Achievement and Teacher Value-Added

Throughout Florida, there is annual testing in Grades 3 through 10 for both math and reading. At the time of our study, two tests were administered: a criterion-referenced exam based on the state curriculum standards known as the FCAT-Sunshine State Standards exam and a norm-referenced test, which is the Stanford Achievement Test (SAT). We employ the SAT in the present analysis because: (a) It is a vertically scaled test, meaning that unit changes in the achievement score should have the same meaning at all points along the scale, and (b) the district under study also administers the SAT in Grades 1 and 2, allowing us compute achievement gains for students in Grades 2 through 10. We use achievement data on the SAT for each of the school years 1999–2000 through 2005–2006.[10] All scores are standardized to the student-level mean of zero and standard deviation of one.

Because the analysis requires having principal assessments and value-added measures for each teacher, we first identified teachers in tested grades and subjects in the 30 schools who had data sufficient to estimate teacher value-added and who were still in the school in the last year for which the administrative data were available, 2004–2005. Many schools had more than 10 teachers meeting the basic requirements for inclusion, and in these cases, we attempted to create an even mix of 5 teachers of reading and math. If there were more than 5 teachers in a specific subject, we chose a random sample of 5 to be included in the list. Even in schools that had 10 teachers on the list based on summer 2005 data, there were cases where some teachers were not still working in the respective schools at the time of the interview (summer 2006). If the principal was familiar with the teacher who had left and felt comfortable making an assessment, then the ratings and comments by the principal were included in the analysis. In six cases where the principal was not sufficiently familiar with the teacher, the teacher was dropped, yielding a total of 294 usable observations.

To obtain the teacher value-added scores, we estimate several value-added measures based on the following general model of student achievement:

Table 1

Descriptive Statistics for Teacher Effectiveness and Characteristics

Teacher Measures	*N*	*M*	*SD*	Minimum	Maximum
Teacher characteristics (raw)					
Intelligent	294	7.93	1.22	2	9
Knows subject	294	7.85	1.33	2	9
Works well with me	294	7.78	1.68	1	9
Communication skills	294	7.62	1.59	2	9
Strong teaching skills	294	7.52	1.60	1	9
Works well with team	294	7.45	1.84	1	9
Caring	294	7.34	1.71	1	9
Motivated	294	7.31	1.82	1	9
Enthusiastic	294	7.20	1.74	1	9
Contributes to school	294	7.01	2.01	1	9
Contributes to community	294	6.95	2.03	1	9
Teacher effectiveness: ratings by principals (raw)					
Math (district wide)	234	7.10	1.68	2	9
Reading (district wide)	231	7.10	1.70	2	9
Teacher effectiveness: ratings by principals					
Math (within school)	234	0.00	1.62	−5.25	2.90
Math (district wide)	234	0.00	1.71	−5.21	1.79
Reading (within school)	231	0.00	1.67	−4.90	2.05
Reading (district wide)	231	0.00	1.80	−5.08	2.64
Teacher effectiveness: value-added (unshrunken)					
Math (within school)	234	−0.061	0.280	−0.907	0.826
Math (district wide)	234	−0.119	0.232	−0.973	0.667
Reading (within school)	231	−0.022	0.282	−0.867	1.509
Reading (district wide)	231	−0.009	0.246	−0.930	1.260

Note. The sample size differs between the characteristics and effectiveness measures because some value-added measures could not be calculated. The number of observations for principal evaluations and value-added are limited to those for whom we have both effectiveness measures. The sum of the observations across subjects exceeds 294 because elementary teachers teach both subjects. In most of the analyses that follow, only the complete observations are used.

$$\Delta A_{it} = \beta_1 X_{it} + \beta_2 P_{-ijmt} + \gamma_i + \delta_k + \phi_m + \gamma_{gt} + \nu_{it}, \quad (1)$$

where X_{it} includes time-varying student characteristics such as student mobility. The vector of peer characteristics, P_{-ijmt} (where the subscript $-i$ is students other than individual i in classroom j), includes both peer characteristics and the number of peers or class size. There are three fixed effects in this base model: a student fixed effect (γ_i), a teacher fixed effect (δ_k), school fixed effect (Φ_m), and grade-by-year (γ_{gt}). The teacher fixed effect captures time-invariant characteristics of teachers. Since school fixed effects

Table 2
Distribution of Teacher Effectiveness Based on Principal Evaluation

	Rating	Elementary	Middle	High	Total
Bottom third	1	0	0	0	0
	2	4	1	2	7
	3	4	2	0	6
		4%	4%	6%	4%
Middle third	4	6	5	5	16
	5	7	12	4	23
	6	27	9	2	38
		22%	33%	31%	26%
Top third	7	46	14	13	73
	8	43	18	6	67
	9	43	17	4	64
		73%	63%	64%	69%
Totals		180	78	36	294

are included, the estimated teacher effects represent the value-added of an individual teacher relative to the average teacher at the school. The final term, v_{it}, is a normally distributed, mean zero error. The model is based on the cumulative achievement model of Todd and Wolpin (2003) and Harris and Sass (2006).

A variety of researchers have questioned the assumptions of this and other value-added models (Rothstein, 2009). Also, while there is some evidence suggesting that teacher value-added estimates are relatively unbiased (Guarino et al., 2010; Kane & Staiger, 2008), there is debate about this (Rothstein, 2009) and there seem to be certain subgroups of teachers for whom the measures are clearly biased (Harris & Anderson, 2012; Jackson, in press). While we wish to recognize the possible concerns here, the relevant point is that these are the measures being used in a growing number of states for teacher accountability and are therefore of interest despite their faults, or perhaps because of them. Also, we are only looking here at *patterns* across teachers. Therefore, even if the measures are biased for individual teachers, this may not introduce bias into our estimates and findings. Recall, for example, that conclusions about the roles of teacher credentials from early studies were relatively unaffected by more recent and elaborate attempts to account for various forms of selection bias.

One criticism of value-added measures is their sensitivity to model specification (Ballou, Mokher, & Cavalluzzo, 2012; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012). We consider six variations in the value-added specification, which vary along three dimensions. First, the

instability noted previously is due to random error in value-added measures. It has now become standard practice to account for this by "shrinking" the estimates so that the value-added estimates of teachers with few student test score observations are pulled back toward the mean (e.g., Rockoff et al., 2010). We also adopted this approach as a sensitivity analysis.

Second, some early work in value-added included the student fixed effects shown in Equation 1, but more recently there has been concern about the level of measurement error and potential biases in this model (Kane & Staiger, 2008; Rothstein, 2009). Therefore, we also estimate models that rely on student covariates (as well as lagged achievement) to account for student differences. Finally, there are advantages and disadvantages to including prior achievement on the left-hand (the gains model) versus right-hand side (partial persistence) of the model and this represents another variation of the model.[11] We estimate six models in all: gains model with unshrunken teacher fixed effects and student fixed effects, gains model with shrunken estimates and student fixed effects, partial persistence model with unshrunken teacher fixed effects and student fixed effects, partial persistence model with shrunken effects and student fixed effects, partial persistence model with unshrunken teacher fixed effects and student covariates, and partial persistence model with shrunken estimates and student covariates.

In math, the correlations among these measures is no lower than +.75, though the correlations are as low as +.50 in reading. To make sure that our results are not sensitive to the value-added model, we therefore re-ranked teachers on the various models and recreate the LVA and HVA groups. In math, 19 teachers of the 47 LVA teachers from the base model were LVA in every specification. In other words, about 40% of the LVA math sample would be the same no matter what value-added model is used. Interestingly, the HVA math teachers are more consistently HVA across specifications with 32 of 47 being in that category in every model. In reading, the numbers of consistent LVA and HVA teachers were 26 and 31, respectively. Put differently, about 60% of teachers at either extreme show no sign at all of being at the other extreme. As we show in the following, the inconsistencies that do exist across specifications do not seem to influence our conclusions.

The descriptive statistics regarding the base value-added measures (Equation 1) are also in Table 1. At the elementary level, all but three of the teachers have value-added measures for both reading and math (no elementary teachers are missing both). We chose the least and most effective teachers separately by subject and carry out separate analyses because the characteristics of effective teachers may vary by subject. There is some missing data on the effectiveness measures because teachers left the schools after we identified them in the prior year's administrative data. Throughout the analysis, we use only the observations with complete data, eliminating the three elementary teachers and 12 secondary teachers who lack any value-added score. The

net result is that for the value-added analysis, we have $n = 234$ in math and $n = 231$ in reading (most secondary teachers taught only one subject). From the original sample of 294 teachers for whom we have usable principal interview data, this yields a total sample of 294 – 15 = 279 with complete data.

Other Methodological Issues

In the previous section, we reviewed some evidence about the validity of value-added and its relationship with principal evaluations. In the following, we consider the validity of our measures from the school principals as well as important issues involved in drawing valid inferences about the relationships between characteristics and effectiveness.

A valid principal effectiveness rating is one that accurately represents, on average, what each principal believes about teacher effectiveness. All indications are that we succeeded in this regard.[12] The situation with the teacher characteristics is somewhat different. Since we cannot validate the teacher characteristic measures, we have to assume that principals' reports of the characteristics are valid measures of those constructs, not just valid indicators of their impressions. But it is worth noting again that principals did seem honest and open in their responses and that we identified the list of characteristics from a combination of prior evidence and open-ended discussions with the same principals in a prior interview, so these are constructs the principals are familiar with and thought about prior to our data collection. Nevertheless, we expect at least some measurement error and bias.

The school average rating may also vary across schools because actual teacher characteristics are not randomly distributed across schools. This means that two teachers in different schools who share the same within-school rating on caring are still different in their true level of caring. Because the nonrandom assignment is a school-level phenomenon, we might think that the solution to the problem of multiple rubrics discussed previously—subtracting the school mean—would solve this problem as well. This is true so long as the nonrandom assignment of teachers is unrelated to the differences in the principals' rubrics. However, if the two problems are interrelated, then it becomes unclear whether subtracting the school mean solves the problem. These are problems inherent to analyzing nonstandardized effectiveness measures across schools.[13] In our judgment, the differences in how principals rate teachers is likely to be greater than the variation in average characteristics of teachers across schools, which implies that subtracting the school mean for the characteristics and effectiveness ratings is preferable to making no adjustments. Therefore, as suggested by the inclusion of the school fixed effect in Equation 1, we use this "within-school" approach throughout the analysis of the relationships between teacher characteristics and effectiveness. We also reanalyze the data without this adjustment and obtain similar answers to our research questions.

The within-school approach plays out somewhat differently with the effectiveness measures compared with the aforementioned discussion of characteristics because we need to separate teachers into low and high categories. As indicated previously, we chose the top two teachers within each school on the principal overall rating to be the "high" rated teachers and the bottom two as the "low" rated, but there were some ties, namely, teachers with the same high or low overall rating. In those cases, we used teachers' average ratings on the personal and professional qualities to identify the teachers with the highest and lowest ratings. (There was no need to break ties with the value-added measures because they are continuous variables.)

Results

Our main question is, why do teacher value-added measures differ from principals' impressions of effectiveness? To begin, we analyze the relationship between the two effectiveness measures. Then, we present our main findings from all the analyses, organized according to both theme and methodology.

The Overlap in Effectiveness Measures Is Modest, but Principals Know the High Flyers

Figure 1 plots the within-school overall principal evaluation and value-added measures for each teacher. The linear relationships are similar—and similarly weak—in both subjects. The correlations are .276 and .168 in math and reading, respectively (significant at $p < .05$). These increase slightly, to .319 and .236, after adjusting for the varying number of students whose scores are available to estimate each teacher's value-added (shrinkage) (Harris & Sass, 2009b). In addition to random error, we show later that these apparently low correlations are also partly due to differences in the construct of effectiveness.

These correlations imply that few teachers are in the same effectiveness category on both measures. Table 3 shows more concretely that only about 30% of the teachers identified as being low (or high) using one measure are also identified in the same category using the other measure. As a basis of comparison, if both measures were of the same construct and involved no measurement error, then the overlap would be 100%, and if the teachers were placed into effectiveness categories at random, the overlap would be about 20%.[14] Therefore, the actual percentage overlap reported in Table 3 is closer to random chance than a perfect relationship.

Looking across all three effectiveness levels (low, middle, high), 139 teachers (59%) are unaffected by the choice of effectiveness measures (i.e., they are in the LVA-LPR, MVA-MPR, or HVA-HPR categories). Conversely, only 10 teachers were at opposite extremes (5 LVA-HPR and 5 HVA-LPR). This might seem to suggest more consistency than the figures

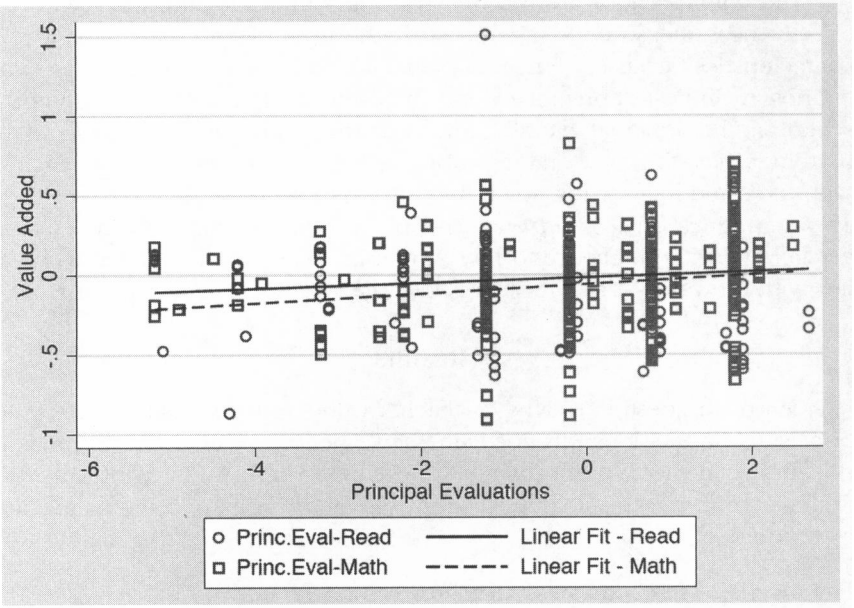

Figure 1. **Relationship between principal evaluations and value-added.**

in the prior paragraph, but this is because we are now considering the middle effectiveness category in addition to the low and high groups; there are more teachers in the middle categories by design, just as there are in current policies that emphasize rewards for small numbers of the highest performers and sanctions for the few with low measured effectiveness. The broader the range of effectiveness included in a category, the greater the consistency between any two effectiveness metrics.

Given the skewness in the overall principal ratings, we might expect principals' ratings to line up less well with value-added measures among the most effective teachers because large numbers of teachers have high ratings. However, consistent with Jacob and Lefgren (2008), the two measures actually line up better at the high end of the effectiveness distribution. Table 3 shows nearly twice the number of teachers ranked high on both measures compared with the number ranked low on both. This pattern arises in both reading and math.

To add further depth, we also examined the interview transcripts and principals' open-ended discussion of each teacher and identified teachers who principals described with superlative terms. These may be more reliable than the numeric scores, given that right skewness of the ratings (see Table 2). There were 27 instances where principals described HPR teachers in both

Table 3
Overlap Between Principal Overall Assessment and Value-Added (Within-School Approach)

	Ratings	Principal Assessment	Value-Added	Teachers Overlapping in the Two Measures
Math				
Elementary	High	36	36	16
	Low	36	36	8
Middle	High	7	7	1
	Low	7	7	3
High	High	4	4	1
	Low	4	4	1
Total		94	94	30
Reading				
Elementary	High	36	36	15
	Low	36	36	7
Middle	High	7	7	2
	Low	7	7	2
High	High	4	4	1
	Low	4	4	2
Total		94	94	29

Note. "Low" and "high" designations based on the bottom 2 and top 2 in the ranked lists of 10 (or sometimes fewer) teachers per school. This differs from Table 2.

reading and math with superlatives such as "the strongest teacher I've got," "exceptional," "outstanding," "cream of the crop," "a super teacher," "excellent in everything she does," and "high flyer." If principal evaluations were unrelated to value-added, we would expect approximately five of these to be HVA teachers (20%). In reality, we found that 70% of the teachers described with superlatives were HVA—the high flyers—and only 11% were LVA (the remaining five are MVA).

The apparent differences between principals' numeric ratings and the superlatives could also be due to the basis of comparison used with the numeric ratings. We therefore compared the overlap in the specific teachers who would be chosen as most and least effective using the within-school *and* whole district approaches for both the value-added and principal evaluation. In additional analysis (available upon request), we show that for both effectiveness measures there is roughly two-thirds overlap in the specific teachers who turn out to be least and most effective using these two alternative methods. As a result, the choice of the method of comparison has little impact on our subsequent findings and we do not discuss it further.

In short, while the correlation between the simple numeric principal ratings and teacher value-added are modest, principals do seem to know who their high flyers are, even if they do not always identify them in the ratings.

Many Characteristics of Effective Teachers Are Consistent Across Effectiveness Measures

The primary purpose of this study is understand why the effectiveness measures differ. Therefore, in the second task, the interviewers asked principals to rate each of the 10 teachers according to the 11 preselected personal and professional characteristics. Table 4 compares the mean characteristics of the least and most effective teachers according to both effectiveness measures. As expected, the means of the characteristic measures for the most effective teachers are almost all greater than the means for the entire sample (shown in Table 1), which in turn are almost always greater than the means of the least effective teachers. The fact that the differences between low- and high-effectiveness teachers are clearer when looking at the principal overall rating is unsurprising given that: (a) this rating and the characteristic ratings come from the same source—the principal and (b) there is a weak relationship between the principal overall rating and value-added indicated in Table 3.

To see what teacher characteristics seem to have the greatest influence on the principals' overall ratings of teacher effectiveness, we subtracted the means of the characteristic measures of the least effective teachers from the means for the most effective teachers, as shown in Table 4, Column 3. Considering the rating by the principal as the measure of effectiveness, the results suggest that the most effective teachers are distinguished by (in order): teaching skills, motivation, enthusiasm, contributions to the community and school, and ability to work well with teams. These differences are all statistically significant. The results are generally similar in math and reading. As shown in Column 6 of Table 4, the same characteristics are also the most important in explaining teacher value-added in math, though motivation is no longer statistically significant. For teacher value-added in reading, contributions to school and community and works well with others become less important and communication and intelligence become more important.

The consistently significant differences in characteristics across the low and high effectiveness groups is partly a function of the high correlations among the characteristics, which range +.5 to +.8. Given this, the differences are likely to be all statistically significant or all insignificant. In developing the interview protocols, we conceptualized some of these measures as reflecting broader latent factors. This is most obvious with measures that have similar names: works well with me and works well with team as well as contributes to school and contributes to community. We also viewed teaching skill, subject knowledge, and intelligence as elements of a single

Table 4

Mean Ratings on Characteristics of Most and Least Effective Teachers (Within-School Approach)

	Principal Evaluations			Value-Added Measures		
Teacher Qualities	(1) HPR	(2) LPR	(3) Difference Between (1) and (2)	(4) HVA	(5) LVA	(6) Difference Between (4) and (5)
Math						
Caring	0.893	−1.213	2.106***	0.127	−0.362	0.489
Strong teaching	1.179	−1.736	2.915***	0.413	−0.311	0.723***
Knows subject	0.865	−1.518	2.383***	0.311	−0.284	0.596***
Enthused	1.023	−1.657	2.681***	0.151	−0.487	0.638**
Motivated	1.145	−1.727	2.872***	0.251	−0.302	0.553
Communication	0.767	−1.531	2.298***	0.001	−0.276	0.277*
Intelligent	0.729	−1.037	1.766***	0.091	0.015	0.076
Works well with team	0.865	−1.646	2.511***	0.312	−0.220	0.532**
Works well with me	0.816	−1.376	2.191***	0.220	−0.120	0.340*
Contributes to school	1.052	−1.480	2.532***	0.158	−0.417	0.574**
Contributes to community	1.033	−1.562	2.596***	0.182	−0.477	0.660***
Reading						
Caring	0.893	−1.064	1.957***	0.170	0.106	0.064
Strong teaching	1.221	−1.800	3.021***	0.476	−0.353	0.830**
Knows subject	0.865	−1.412	2.277***	0.269	−0.178	0.447*
Enthused	1.045	−1.487	2.532***	0.364	0.002	0.362*
Motivated	1.060	−1.685	2.745***	0.379	0.145	0.234
Communication	0.830	−1.467	2.298***	0.235	−0.148	0.383*
Intelligent	0.729	−1.037	1.766***	0.176	−0.249	0.426**
Works well with team	0.907	−1.497	2.404***	0.078	0.035	0.043
Works well with me	0.816	−1.333	2.149***	0.135	0.114	0.021
Contributes to school	1.115	−1.438	2.553***	0.201	0.222	−0.021
Contributes to community	1.076	−1.562	2.638***	0.225	0.097	0.127

Note. Effectiveness categories are the same as in Table 3. HPR = high rating by the principal; LPR = low rating by the principal; HVA = high value-added; LVA = low value-added. Chi-square test of statistical significance.
*p = .10. **p = .05. ***p = .01.

construct, technical skill. We therefore carried out a factor analysis to attempt to identify the latent constructs.

As the theoretical structure of the latent characteristics is not well established in the literature, we conducted exploratory rather than confirmatory factor analysis, using maximum likelihood and principal factors routines in Stata. Given the high correlations among the measures, we expected positive

correlations among the factors and therefore used oblique rather than orthogonal rotation. To identify the appropriate number of factors, we follow Onwuegbuzie et al. (2007) and use a combination of theory, interpretability (i.e., whether the factors relate to well-defined constructs), and the screeplot method (Zwick & Velicer, 1986). The K1 eigenvalue method (Kaiser, 1958) is also common, although this is best viewed as establishing a lower bound on the number of factors. The K1 approach suggested at least two factors and the screeplot suggested two to four factors. The resulting factor loadings yielded four easily interpretable factors that lined up closely with our theoretical framework: technical skill, affective traits, team orientation, and contributions outside the classroom. The factor loadings for these are available in the appendix available from the authors upon request. In the analysis that follows, we extend our analysis of the individual measures and present new analysis of the four factors. The use of the factors also has the side benefit of reducing the multiple comparisons problem, namely, that testing for differences among more variables increases the probability of finding at least one statistically significant correlation by chance alone. Using the four factors reduces the number of comparisons considerably.

Table 5 provides results of regressions of the two effectiveness measures on both the individual characteristic measures and the four latent factors (errors clustered at the principal level) to identify the strongest predictors. As shown in Table 5, Column 2, principal evaluations for teachers in both reading and math continue to be positively correlated with teaching skill, communication, and motivation. Knowledge of subject seems to be important in math, but not reading, while working well with the principal is associated with principal evaluations in reading.

The equivalent results for the value-added measures, shown in Column 5 of Table 5, are no longer statistically significant, except for teaching skill, subject knowledge, and intelligence among reading teachers. The limited statistical significance is unsurprising given the low reliability of value-added measures, the high correlations among the covariates, and the modest sample size. The inconsistency in value-added measures across specifications noted previously also led us to conduct robustnesss checks. The results reported previously are based on Equation 1, what we call the base model, and we compared this to the results when using value-added measures from an "alternative" value-added model: shrunken estimates with partial achievement persistence and student demographics instead of student fixed effects. This yields the sharpest possible contrast with the base model, which does not use shrinkage, assumes complete persistence, and relies on student fixed effects.

Under the alternative model, as shown in Column 6, the math results remain insignificant, except for working well with the principal, which is negatively associated with value-added. Teaching skills and intelligence continue to predict teacher reading value-added. Several other characteristics

Table 5

Coefficients From Regressions of Teacher Effectiveness Measures on Teacher Qualities and Latent Factors

	Dependent Variable					
	Principal Evaluations		Value-Added			
Teacher Qualities	(1)	(2)	(3) Base	(4) Alternative	(5) Base	(6) Alternative
Math						
Affect traits	0.547***		0.016	0.015**		
Caring		–0.036			–0.018	–0.010
Enthused		0.078			0.007	0.017
Motivated		0.178**			0.013	0.001
Technical skill	1.100***		0.066***	0.048***		
Strong Teaching		0.485***			0.032	0.005
Knows Subject		0.187*			0.005	0.021
Intelligent		0.033			0.021	0.025
Communication		0.166***			–0.006	–0.015
Team player	0.230**		–0.023	–0.014		
Works well with team		0.085			0.011	0.010
Works well with me		–0.057			–0.025	–0.019**
Contributions beyond class	0.242*		–0.002	0.002		
Contributes to school		–0.015			0.004	–0.013
Contributes to community		0.050			–0.008	0.015
N	234	234	234	234	234	234
R^2	0.736	0.755	0.050	0.085	0.056	0.104
F-test p-value	0.000	0.000	0.007	0.000	0.089	0.000

(continued)

Table 5 **(continued)**

	Dependent Variable					
	Principal Evaluations		Value-Added			
Teacher Qualities	(1)	(2)	(3) Base	(4) Alternative	(5) Base	(6) Alternative
Reading						
Affect traits	0.403***		–0.005	0.004		
Caring		–0.020			–0.024	–0.012**
Enthused		–0.170**			0.024	0.012
Motivated		0.235***			–0.026	–0.003
Technical skill	1.073***		0.046**	0.035***		
Strong teaching		0.676***			0.062***	0.023***
Knows subject		–0.082			–0.049*	–0.003
Intelligent		0.090			0.044*	0.013*
Communication		0.160***			–0.020	–0.002
Team player	0.277***		–0.020	–0.003		
Works well with team		–0.090			0.002	0.004
Works well with me		0.183**			–0.006	–0.006
Contributions beyond class	0.366***		0.011	–0.012*		
Contributes to school		0.081			–0.010	–0.017***
Contributes to community		0.057			0.021	0.009*
N	231	231	231	231	231	231
R^2	0.698	0.722	0.037	0.128	0.056	0.169
F-test p-value	0.000	0.000	0.031	0.000	0.000	0.000

Note. Each column is a separate regression where the dependent variable is the indicated effectiveness measure (within-school approach). As a robustness check, we report results for the "base" value-added model shown in Equation 1 as well as an "alternative" model: shrunken estimates with partial persistence and student covariates instead of student fixed effects. Statistical significance based on clustered standard errors reported. The factors are listed first (see appendix available from the authors upon request), and individual components are indented
*p = .10. **p = .05. ***p = .01.

become significant, but given the small sample, multiple comparisons, and number of methodological variations, we emphasize only the results that are consistent across the two models. In that regard, teaching skill and intelligence consistently stand out.

When we reanalyze the data using the four factors, the results are more consistent across specifications and compared with the simpler difference-in-means tests in Table 4. All four factors are positively correlated with the principal evaluation with p-values of .05 or better in both subjects. Technical skill still stands out as the strongest predictor, as we would expect given that teaching skill is heavily loaded on to this factor. Technical skill is also the only positive and statistically significant predictor of teacher value-added in both subjects. This suggests that the sporadic statistical significance of the individual predictors from the nontechnical factors (e.g., works well with me) are probably misleading.

The relatively weak value for teacher intelligence among principals in Column 2 is consistent with Ballou (1996), who concludes that intelligent teachers do not appear to be given high ratings by principals, but only partly consistent with his assumption that the most intelligent teachers are generally "best" in terms of generating academic learning. Intelligence does seem to predict teacher value-added in reading, but not in math. One reason for the weak relationship between effectiveness and intelligence may be that this is the characteristic that had the highest average rating, resulting in less total variation across teachers. This is certainly not the only explanation, however, as the characteristic with the second highest average rating—subject knowledge—is associated with teacher effectiveness in all the analyses.

These results suggest that the relative importance of the prespecified characteristics is similar across the two effectiveness measures, though our more in-depth analyses of the interview data suggest noteworthy differences.

A Deeper Look: How Principals Prioritize Effort

In this section, we draw on the principals' open-ended comments to understand how the principals described teachers across the HPR, LPR, HVA, LVA categories. We focus on characteristics not considered in the aforementioned quantitative analysis, specifically those that emerged in principals' descriptions of these teachers. In this analysis, we indicate the proportions of teachers reporting particular responses that we had coded and provide quotations that highlight these points with greater depth and clarity, though we do not provide statistical tests as this might be construed as creating a false sense of precision in the coding of the variables themselves.

In the HPR-LPR comparison, three characteristics emerge as being most important to principals: professional development, experience and burnout, and family and personal situations. We discuss each in turn.

Professional Development

Principals described 24 of 60 HPR teachers as willing to seek professional development, but only 2 of the 60 LPR teachers were described in this way. Common in the descriptions of HPR teachers were phrases such as "I would consider [him] over and above for professional development. He's always trying to extend his knowledge" and "She's constantly going after learning new methods and learning how to do something better." Principal T described a highly rated teacher as one who "stays up on the subject matter; that is always searching for the best teaching skills." Similarly, 7 highly rated teachers were noted as having obtained National Board for Professional Teaching Standards (NBPTS) certification. We interpret NBPTS as a form of professional development for purposes here because it requires 200 to 400 hours of work and training (Goldhaber & Anthony, 2007). No LPR teachers were identified as having obtained or pursued NBPTS.

Principals complained that 5 of 60 LPR teachers were not proactive regarding professional development. For example, regarding one LPR teacher, a principal said, "She doesn't want to volunteer for any in-services any time, you know. If she's done something once, well, then she feels she knows everything." Principal Y stated, "Even if you've been around 20, 30 years, or whatever, you can always still learn." No HPR teacher was described as unwilling to pursue professional development.

One possible interpretation of this is that principals value teachers who try to improve, perhaps regardless of how much success they actually have in raising student test scores. Indeed, while principals clearly value professional development, there is little evidence to suggest that these efforts generally pay off in terms of higher student test scores (Garet et al., 2008, 2010; Harris & Sass, 2011; Jacob & Lefgren, 2004). Principals value teachers who keep up with new curricular and instructional practices and recognize the time and effort by Nationally Board Certified teachers, but again, while National Board teachers have somewhat higher value-added, this is mainly a result of selection rather than improvement occurring as a result of the extensive NBPTS certification process (Goldhaber & Anthony, 2007; Harris & Sass, 2009a).[15]

Experience and Burnout

Another contrast between the way that principals talked about HPR teachers and LPR teachers is years of experience. For 11 of the 60 HPR teachers, being a "veteran" was discussed in a positive light, for example, the teacher "has taught generations of children" or "has done it for so long, she knows what works and what doesn't." But principals also identified some highly rated teachers as less experienced. For example Principal B described one HPR teacher (who was also HVA) with 4 years of experience as "a shining star." Similarly, Principal V described one of his "high flyers" as

being "a couple years out of school . . . and she loves the kids. She's a strong teacher." Principals acknowledge that despite their relative inexperience, these beginning teachers are skilled and already proving themselves capable and productive.

For 15 of the 60 LPR teachers, however, experience was raised in a negative light. In 11 of 15 LPR cases, principals noted that teachers suffered from burnout. One LPR was described as "a bit older, a bit worn, a bit tired." Burnout was also raised with two HPR teachers. The fact that experience seems to cut both ways—increasing skill in some but decreasing motivation and enthusiasm in others—corroborates findings from the earlier first round of interviews with principals in this same school district, which focused on principals' views about teachers in general rather than specific teachers in their schools (Harris et al., 2010). It is important to emphasize, however, that experience came up far less often in our interviews than professional development, discussed earlier. This, as well as the next theme, should therefore be viewed as more exploratory.

Family and Personal Situations

Principals' discussions of professional development and burnout are not the only ones that point toward the importance of effort.

Principals identified 11 of 60 LPR teachers as dealing with personal situations such as divorces, deaths in the family, or serious illness that explained lower ratings. For example, Principal V described an LPR teacher stating that she has a "personal life that's pretty consuming—a divorce and kids." Principal P simply described an LPR teacher as "having a lot on her plate from a personal standpoint that limits the amount of time that she can do additional things."

Like the discussion of professional development, the fact that principals volunteered information about teachers' personal situations suggests the importance principals place on the time, effort, and focus put forth by teachers. Given that many teachers do not experience the kinds of personal obligations and difficulties that principals mentioned, and that many others would try to keep these issues to themselves, the fact that personal situations came up in 15% of the cases seems noteworthy. Personal issues came up for a smaller number of HPR teachers (6 of 60), but principals explained that despite these circumstances, they were still able to perform well as teachers.

The Effort Paradox

The focus on professional development, burnout, and family issues together suggest a larger theme about the importance that principals place on effort. On one level, the fact that principals focus on effort is understandable and predictable. It is difficult to imagine any leader, manager, or supervisor not wanting people to work hard. Effort is also readily observed.

Principals can notice teacher effort when they arrive early or stay late at school or put in time leading committees. Effectiveness, however, is harder to see, especially with the traditional brand of evaluations where principals spend little time observing teachers in the classroom.[16] Further, while some forms of effort, such as professional development, are aimed at improving teachers' instruction, principals cannot readily determine whether this improvement is occurring.

The same logic has been found to extend to how teachers evaluate their students. Cross and Frary (1999) suggest that teachers may assess students with less "technical purity" and rather than focusing solely on students' observed "performance," grade students based on growth and improvement, conduct, attitude, potential ability—and effort. Cross and Frary conclude that even when trained in recommended measurement and assessment practices that focus on knowledge and skills relevant to a course, teachers still focus their grading on these other criteria. Principals appear to take the same approach with their teachers.

Some principals, though, seem to take this a step further and assess teachers not only on effort per se, but effort that leads to changing practices. Principal I described one HVA (reading) teacher this way:

> She cares for the kids. I would say, pretty good there, but her skills in the classroom are not good and her subject area is very weak. She doesn't involve herself with her department members, and she doesn't do a lot of the staff development type things and pretty much does the same thing she's done probably for all her years as a teacher. I have a lot of teachers that teach, and they teach 30 years, and they teach the same lessons 30 times. She does work well with me. I marked her pretty decently with better than adequate as far as her doing things other than just classroom stuff; but quite frankly, she's probably one of the few that I'd rather keep in the classroom and a little bit less of the afternoon stuff.

Notice that the primary basis for saying that "her skills in the classroom are not good" is that she "pretty much does the same thing she's done probably for all her years a teacher." Is it possible that the teacher has a method that really works for her and her students, so that doing the "same thing" actually make sense? This is not a question we can answer here, nor is it one that seems to have occurred to this principal.

These results lead to a bit of a paradox. While there are circumstances and strong norms driving principals to focus on teacher effort—the same pressures affecting teachers' evaluations of students—the kinds of effort that principals prioritize may not be good proxies for actions that increase student learning and test scores. If principals only notice efforts that lead to changes in instruction as opposed to approaches that work well to improve student achievement, they may inadvertently reward the displacement of effective practices. Likewise, if they reward ineffectual professional

development, they may simply take time away from the more important tasks of instruction.

High Value-Added Teachers as Lone Wolves

The types of teachers rewarded by value-added measures may also differ from principal evaluations because of teachers' contributions to the school and community. Unlike the rest of the prespecified list, the connection of these outside-the-classroom contributions to meaningful positive influence on test scores of teachers' own students is more tenuous (except insofar as these activities involve direct discussions about teachers' own instruction). Principals, on the other hand, have schools to run, student extracurriculars and activities to organize, and faculty committees to manage—and a desire to facilitate a collective efficacy in meeting common objectives. They cannot do all this themselves and therefore are likely to value teachers who contribute to the larger organizational effort.

While highly effective teachers generally have more positive characteristic ratings with both the value-added and principal evaluations, teacher contributions to schools and community do seem to play a large role in principal evaluations for reading teachers. Contribution to the school ranks last in Table 4 as a factor distinguishing LVA and HVA reading teachers but ranks fourth in importance when comparing LPR and HPR teachers. The fact that this same pattern does not appear in math may be because principals see math as more technically challenging and therefore weight intelligence and other factors as more important. Contributions to school and community also seem less important to principals in elementary schools, presumably because these schools have fewer students and teachers for principals to deal with and fewer extracurricular activities (results available upon request).

Three principals described some HVA teachers as isolating themselves in the classroom—what one principal called "lone wolves"—and this too seems to reflect the importance principals give to contributions outside the classroom. Principal D described one such teacher who had high value-added in reading:

> She's just a very quiet, stay-in-her-room kind of person. Excellent communication. She works fairly well with her grade and team. I would say the reason I marked her at the bottom of exceptional [7 on a 9-point scale] is she doesn't always do her part. She doesn't always show up to a team meeting, and I'm not sure she always carries her load as much as she possibly could.

The lone wolf scenario is also reinforced by the results in Table 5, showing that principal ratings on works well with me are consistently negatively associated with value-added (although only statistically significant in one of four cases). Nevertheless, many HVA teachers are not lone wolves. We found that

among the HVA teachers there were seven instances of principals saying that HVA teachers were good mentors for other teachers, compared with only two such cases for LVA teachers. Different principals may view this in different ways.

Philosophical and Personality Divides

Principals identified two aspects of teachers' instructional approaches that were associated with their assessments of teachers. Among HVA teachers, 5 of 72 were described as teachers who "set high standards for students," where only 1 of the LVA out of 72 was described in this way. Furthermore, 7 of 72 HVA teachers "provide an active learning environment," compared with only 3 of 72 LVA math teachers. These patterns held only at the elementary level.

One example is worth highlighting because it suggests that some teachers might be given low ratings by principals because of disagreement over instructional philosophy. Principal BB said the following about one HVA teacher:

> This is a teacher who's not coming back this year. She chose not to, but I think she chose not to sort of because of me. I am really, really into student accountability, and, you know, I just felt like she wasn't. I think she'd been teaching for about ten years. I think she cared. The other [former] principal here, she may have had a real strong bond to, but I felt like she and I didn't bond. I felt like she never really would come into my circle or believe in my philosophy, and I think she had the feeling that some kids can't learn, and I don't have that philosophy. . . . Her scores were okay, but was one who didn't believe that you need to put a lot of emphasis on testing.

This last sentence is particularly telling. Her students were apparently doing well on standardized test scores, suggesting the principal realized this was an HVA teacher, but the principal perceived a difference in philosophy that led to conflict—and a low rating of 5 (sample average of 7.1).

One potential explanation is that the principal viewed "student accountability" as more than high test scores, focusing more on what teachers communicated about their beliefs and instructional perspectives rather than bottom-line results. More likely, however, is that what we have described so far as a philosophical divide was really about personality and perceived loyalty. The focus of this principal's comment on the teacher's relationship with the prior principal suggests a possible power struggle or a poor relationship with the teacher, which may have shaped his or her view about the teacher's beliefs and skills. The principal talks about "having the feeling" that the teacher thought "some kids can't learn," but did not refer to specific statements or actions taken by the teacher. This principal may be jumping to conclusions, which itself has implications for the role of school principals in

evaluating teachers. The importance of instructional philosophical and personality are reinforced by other research on this same group of principals suggesting that principals hire teachers who "match" the philosophy and culture of their schools (Harris, Rutledge, Ingle, & Thompson, 2010). While this is only one example, and the value-added measure might just be incorrect for this teacher, it illustrates the complexity and subjectivity inherent in administrator-teacher relations and raises important questions about what aspects of teaching are captured by different effectiveness measures.

A Conflict Between Caring and Test Scores?

We predicted that the LVA-LPR would have the lowest characteristics ratings and that HVA-HPR would have the highest characteristic ratings, but this turns out not to be the case. The five HVA-LPR teachers have the lowest characteristics ratings of any group on average, while the MVA-HPR teachers have the highest. Specifically, MVA-HPR teachers were rated higher in caring and motivation. This could reflect a perception by principals that teachers who focused heavily on the bottom line of student test scores (generating the high value-added) are automatically less caring. For example, Principal H described one HVA-HPR teacher this way that reveals the potential tension:

> Caring is less strong than the teaching skills. Teaching skills are some of the strongest I've seen. Strongest among this group. She knows her subject areas backwards and forwards and takes training, attends workshops. . . . She is enthusiastic and generally keeps a good attitude, but with certain kids and certain things that have happened that detract from her original enthusiasm. Motivation is pretty strong. She definitely is motivated to improve her test scores.

This principal describes the teacher as hard driving and test focused (not to mention, again, being focused on professional development workshops). The principal still gives the teacher a high rating, but with a significant caveat that she is less enthused about "certain kids," something the closed-ended response could not have revealed.

This philosophical divide reinforces the importance, as well as the complexity, of the construct of teacher effectiveness. Principals' views of teachers are clearly affected by teacher' philosophies, personalities, loyalties, and attitudes, regardless of how this translates into classroom instruction.

Discussion and Limitations

Why do the effectiveness measures differ? At first glance, it does not appear that there is much difference in the underlying construct of effectiveness because the characteristics associated with each effectiveness measures

are similar. But our mixed-methods approach reveals that this is misleading. Teachers give higher evaluations to students based on their effort, and principals seem to do the same with their teachers. This pattern emerges from several different directions—explicit references to effort as well as indirect references to challenging family situations and professional development.

Whether these differences in the associated characteristics reflect divergent constructs of effectiveness is harder to determine. While there is little evidence that formal professional development has much influence on teacher value-added, it would be reasonable for them to connect teaching skills with academic learning. But principals also frequently discussed affective traits, and some were critical of teachers who seemed uncaring in their pursuit of academic excellence. In discussing teachers' contributions to the school as a whole, they rarely if ever suggested that non-classroom activities were important for raising scores, and of course many school activities, such as sports and social events, do not have academic achievement as an immediate objective. Principals, as well as teachers, parents, and students, want their schools to be proper learning environments, but on some level they are mainly trying to keep schools running smoothly. Our qualitative findings reveal potential tensions between principals' organizational and instructional goals.

As further evidence, recall that we asked principals to rate teachers overall as well as in their contributions to student test scores. These two metrics are correlated at 0.733 in math and 0.741 in reading. If principals had seen these two as one in the same, these correlations would have been closer to 1.0. They were asked about test score contributions after the overall rating and they could have repeated their answers. These principals clearly distinguish between contributions to test scores and other contributions.

Given that principals are likely to play a significant role in actual high-stakes evaluations, and that their perspectives will play a role no matter how standardized the observation rubric, understanding principals' views about teacher effectiveness is a critical issue. The largest of the current efforts to understand various effectiveness measures—the Gates Foundation's MET project—includes classroom evaluations carried out by highly trained observers, rather than more realistic evaluations by actual principals and other observers that give considerable weight to what occurs behind the classroom door, but also to a larger concept of organizational contributions.[17] Similarly, the evaluation rubrics used in MET, such as the Danielson Framework, focus almost entirely inside the classroom and give very little attention to activities outside the classroom.[18]

We cannot say how far these conclusions extend beyond our sample of 30 principals and 294 teachers in this study. The depth of our analysis comes with some sacrifice in breadth and representativeness. Also, as we pointed out earlier, both sets of effectiveness measures are low stakes; while student test scores are important and the participants in the study work within Florida's high-stakes accountability environment, no personnel decisions

were based on the measures we collected at the time we collected them. To the degree that Campbell's Law operates, the more recent increase in the stakes may alter the relationships observed. Principals themselves are also increasingly being held accountable, and this could very well change what they look for in teachers.

Conclusions and Policy Implications

In the new era of Race to the Top, and the eventual reauthorization of the federal Elementary and Secondary Education Act (ESEA), state and local policymakers face an important question: How should we use alternative measures of teacher effectiveness such as value-added to evaluate teachers and hold them accountable? We cannot answer this for them, but our results do have much to say about the implications of their decisions.

State and federal policymakers should recognize that the various measures differ not just in their validity, but in the construct they measure—valid measures of what? Principals, even in a strong accountability state such as Florida, do not focus solely on test scores when identifying teachers' characteristics of effectiveness. They run schools—schools that have complex and often competing missions and a need for collegiality—while being subject to forces from multiple external stakeholders and many levels of government. A role for principals in evaluation also seems warranted because they have a great deal of information about their teachers, from parent requests and inquiries, students, other teachers, and their own direct observations. As schools are organizations that are formally led by principals, it seems essential for principals to have a say.

Ironically, as more demands are placed on principals to evaluate teachers, principals may be forced to lean on their teachers to perform other important duties—the same outside-the-classroom activities that policymakers are, intentionally or not, pressuring principals to downplay. More aggressive accountability may also change the nature of teacher-principal relationships and induce principals to pay less attention to the effort and isolating behavior of the lone-wolf teachers that predominated in these interviews.

Policy decisions on these matters will have consequences. We, as well as others (Harris & Sass, 2009b; Jacob & Lefgren, 2008; Rockoff et al., 2010), show that different teachers will be identified as effective by value-added measures versus principal evaluations. However, in a rewards-oriented system, where only high effectiveness designations are relevant, the decision may be less consequential than it appears. Principals know who their high flyers are. Interestingly, much of the debate has focused instead on dismissing low-performing teachers rather than rewarding and retaining the high-performing ones (Gordon, Kane, & Staiger, 2006; Hanushek & Rivkin, 2010). Whatever the other merits of this approach, our results suggest that "incorrect" employment decisions from the standpoint of student achievement

are more likely to emerge among less effective teachers who may have philosophical or personality-driven conflicts with principals.

To the degree that principals are given a role, it appears that principals might be well served by reconsidering how they value different forms of effort. While an ethos of effort is noteworthy for any organization, effort of the sort we heard about from these principals may contribute little to either organizational culture or student learning. It is not that principals should cease from expecting teachers to contribute to the school and community outside their own classrooms. In fact, this might serve as a useful counterbalance to the focus of value-added on classroom contributions, but principals may need to reconsider the value of lone wolves who, even with their apparent obstinacy, do the same thing year in and year out—but do it well.

The consequences of these decisions extend to the types of teachers who will be rewarded. There are important similarities in the characteristics associated with value-added and principal evaluations (e.g., teaching skills are important in both cases; see Table 4), as well as important differences (e.g., the role of professional development and effort more generally). If the current trend toward aggressive teacher accountability becomes institutionalized, then over time this will influence the types of people who are attracted into teaching and the characteristics of those who choose to make it a career. How we evaluate teachers will likely affect the character of the learning environment and the teachers and teaching that students experience.

Notes

We thank Cynthia Thompson for excellent research assistance and David Monk and Robert Floden for their valuable comments. The authors are grateful for generous funding from the United States Department of Education (grant R305M040121), a joint project with Tim R. Sass. The authors are responsible for all remaining errors.

[1]These studies all use longitudinal student achievement data, which we define within the value-added category even though the specific approaches to analyzing these data have changed in recent years. See also Armor et al. (1976) who used snapshot of student achievement rather than longitudinal data.

[2]Specifically, the maximum correlation between any two measures is the square root of the product of the two reliability coefficients. The data used in the present study are insufficient to estimate reliabilities. See the Measures of Effective Teaching (MET) project studies for evidence on reliability of classroom observations.

[3]The national data on principals comes from the 2003–2004 Schools and Staffing Survey (SASS) as reported in the *Digest of Education Statistics 2006* (Snyder, Dillow, & Hoffman, 2007). Part of the reason that this sample of principals has higher levels of educational attainment is that Florida law makes it difficult to become a principal without a master's degree.

[4]We analyzed demographic characteristics of students and teachers within the sampled district, state, and nation in 2004. Data are provided by the Florida Department of Education (2005) and the *Digest of Education Statistics 2006* (Snyder et al., 2007).

[5]As part of the interview, we discovered that principals have two ways to access student test scores that might allow them to evaluate individual teachers. First, many made

use of a district-purchased software program, Snapshot. Second, the district provides state-determined measures to the principals. After inquiring with district officials, we found that both sources of information only calculate mean achievement gains of each teacher's students. As indicated earlier, this does not qualify as value-added per se, but these measures are correlated with value-added. While we have no data about the actual usage of either source of information, the open-ended responses by principals in the formal interviews, as well as subsequent informal conversations with two principals, suggest that at least some principals used the program to look at the achievement gains made by students of each teacher. This likely influenced their responses to some of the interview questions.

[6]For details about the interviews, see Harris, Rutledge, Ingle, & Thompson (2010); Rutledge, Harris, & Ingle (2010); and Rutledge, Harris, Thompson, & Ingle (2008).

[7]To ensure confidentiality of the teachers, the interviewers had a sheet with a list of non-identifiable numbers created by the district. The interviewers were given the lists with the names in sealed envelopes with signatures signed over the seals. The interviewer brought the respective envelope to each interview and handed it to the principal who then opened it. The interviewers asked about the specific teachers using the numbers and the principals used their list to determine the correct name. After the interview, the principals were advised to discard the list.

[8]Here, we provide examples as illustrations of our coding process. A principal discussed a teacher's subject matter knowledge, stating, "He knew his subject and knew it well," which was coded as Subject Matter Knowledge-Positive. Another principal described a teacher, stating, "She cares for the kids. I would say, pretty good there, but her skills in the classroom are not good and her subject area is very weak," which was coded as Caring-Average, Teaching Skills-Negative, and Subject Matter Knowledge-Negative. A similar process was used for each and every other descriptor of individual teachers used by the principals. For example, a principal described a teacher as "He knew the kids, was here after hours, was willing to go that extra mile to help a child after school if they didn't understand and did not ask for extra pay or comp time," which was coded as Gets to Know the Child-Positive, Before and After Hours-Positive, Goes Above and Beyond-Positive.

[9]One possible reason why we did not find teachers with lower ratings is that in our selection of the 10 teachers from each school, we identified only those with value-added scores. Since these scores are available only after 2 or 3 years, weak teachers might have been released earlier.

[10]Prior to 2004–2005, Version 9 of the Stanford Achievement Test (SAT-9) was administered. In 2004–2005 the SAT-10 was given. All SAT-10 scores have been converted to SAT-9 equivalent scores.

[11]In the partial persistence model, the influence of prior achievement is flexible and emerges from the estimation, but the coefficient is likely to be biased due to correlation in the error of A_{it} and $A_{i,t-1}$. In the gains model, we avoid the error correlation but impose a possibly false restriction that prior achievement persists completely (i.e., the coefficient on prior achievement is one).

[12]While we believe we effectively dealt with concerns about confidentiality, a related concern is that principals might just tell the interviewers "what they think they want to hear." For example, they might have thought that the interviewers wanted to see that teachers rated as high overall also had "strong teaching skills." To address this, the interview was designed to separate the questions about the overall ratings by principals from the specific characteristic ratings, so that the principals would be less likely to think that the interviewers were interested in the relationships among the measures.

[13]To provide some sense of the differences in ratings across schools, we tested whether each school's mean rating was different from the district average. Taking caring as an example, we find that the equivalence of the school and district means can only be rejected at the .10 level for 5 of 30 schools. While this in no way proves anything about the differences in principal rubrics and mean teacher characteristics, it does suggest that the influence of these methodological issues might be small.

[14]Since we are selecting the bottom and top 2 out of 10 teachers in each school, there is a 20% chance that the teachers in the low (high) category in the first round of random selection is also in the low (high) category in the second round.

[15]As additional analysis, we compared high value-added (HVA) and low value-added (LVA) teachers based on what principals said about their professional development efforts. LVA teachers in both subjects were somewhat more likely than HVA teachers to be identified as pursuing professional development (reading: 7 vs. 3; math 5 vs. 2). Even if we believe these slight differences represented real patterns, this does not necessarily mean that professional development is ineffective, since low-performing teachers might be more likely to obtain professional development in order to improve on their low effectiveness. Nevertheless, these patterns are consistent with prior research.

[16]Our data were collected well before the changes in teacher evaluations precipitated by the federal Race to the Top.

[17]In the MET project, one of the main objectives is to identify the most valid rubrics for evaluating classroom teaching (Gates Foundation, n.d.). This is being accomplished by testing which rubrics yield measures that are most highly correlated with the value-added measures, implying that value-added is the most accurate measure of teacher effectiveness.

[18]The Danielson framework does include "participating in professional community," but it is only 1 of 22 evaluation factors that focus almost entirely on instruction (Danielson, 2013).

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in Chicago Public Schools. *Journal of Labor Economics, 25*:95–135.

Abernathy, T., Forsyth, A., & Mitchell, J. (2001). The bridge from student to teacher: What principals, teacher education faculty and students value in a teaching applicant. *Teacher Education Quarterly, 28*, 109–119.

Armor, D., Conry-Oseguera, P., Cox, M., King, N. J., McDonnell, L. M., Pascal, A. H., . . . Zellman, G. L. (1976). *Analysis of the school preferred reading program in selected Los Angeles minority schools.* Santa Monica, CA: Rand Corporation.

Baker, B. D., & Cooper, B.S. (2005). Do principals with stronger academic backgrounds hire better teachers? Policy implications for improving high poverty schools. *Education Administration Quarterly, 41*, 413–448.

Ballou, D. (1996). Do public schools hire the best applicants? *The Quarterly Journal of Economics, 111*, 97–133.

Ballou, D., Mokher, C. G., & Cavalluzzo, L. (2012). *Using value-added assessment for personnel decisions: How omitted variables and model specification influence teacher's outcomes.* Unpublished manuscript.

Bidwell, C. E. (2001). Schools as organizations: Long-term permanence and short-term change. *Sociology of Education, 74*, 100–114.

Borghans, L., ter Weel, B., & Weinberg, B. (2008). Interpersonal styles and labor market outcomes. *Journal of Human Resources, 43*, 815–858.

Braun, J., Willems, A., Brown, M., & Green, K. (1987). A survey of hiring practices in selected school districts. *Journal of Teacher Education, 38*, 45–49.

Broberg, J. P. (1987). *Ranking criteria for hiring newly certified teachers: A delphi technique* (Unpublished doctoral dissertation). Oklahoma State University, Stillwater, OK.

Cain-Caston, M. (1999). A survey of opinions of North Carolina school administrators regarding factors considered most important in hiring teachers for their first teaching positions. *Journal of Instructional Psychology, 26*, 69–73.

Campbell, D. T. (1976). *Assessing the impact of planned social change.* Retrieved from ERIC database (ED303512).

Carnoy, M., & Loeb, S. (2003). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*, 305–331.

Clotfelter, C., Ladd, H., & Vigdor, J. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources, 45*, 655–681.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*, 53–72.

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica, 78*, 883–931.

Danielson, C. (2013). *The framework for teaching: Evaluation instrument.* Downloaded from http://www.danielsongroup.org/userfiles/files/downloads/2013EvaluationInstrument.pdf.

Darling-Hammond, L., Amrein-Beardsley, L., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation: Popular modes of evaluating teachers are fraught with inaccuracies and inconsistencies, but the field has identified better approaches. *Phi Delta Kappan, 93*(6), 8–15.

Denzin, N. K., & Lincoln, Y. S. (1998). *Collecting and interpreting qualitative materials.* Thousand Oaks, CA: Sage.

Dunton, J. (2001). *Selection criteria used by high school principals in Virginia when hiring first year career and technical teachers* (Unpublished doctoral dissertation). Virginia Polytechnic Institute and State University, Blacksburg, VA.

Elmore, R. F. (2000). *Building a new structure for school leadership.* Washington, DC: The Albert Shanker Institute.

Fenstermacher, G. D., & Richardson, V. (2007). On making determinations of quality in teaching. *Teachers College Record, 107*, 186–213.

Florida Department of Education. (2005). *Minority representation of Florida's public school teachers, fall 2004.* Tallahassee, FL: Author.

Gallagher, H. A. (2004). Vaughan Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement. *Peabody Journal of Education, 79*, 79–107.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., . . . Sztejnberg, L. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4031). Washington, DC: U.S. Department of Education.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Walters, K., Song, M., . . . Doolittle, F. (2010). *Middle school mathematics professional development impact study: Findings after the first year of implementation* (NCEE 2010-4009). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gates Foundation. (n.d.). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study.* Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *Review of Economics and Statistics, 89*, 134–150.

Gordon, R., Kane, T., & Staiger, O. (2006). *Identifying effective teachers using performance on the job.* Washington, DC: The Brookings Institution.

Grissom, J., & Loeb, S. (2011). Triangulating principal effectiveness: How perspectives of parents, teachers, and assistant principals identify the central importance of managerial skills. *American Education Research Journal, 48*, 1091–1123.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. (2010). *Evaluating value-added models for estimating teacher effects.* Unpublished manuscript.

Hallinger, P. (2005). Instructional leadership and the school principal: A passing fancy that refuses to fade away. *Leadership and Policy in Schools, 4,* 221-239.

Hallinger, P., & Heck, R.H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980–1995. *Educational Administration Quarterly, 32,* 5–44.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30,* 466–479.

Hanushek, E. A., & Rivkin, S. G. (2010). *Using value-added measures of teacher quality.* Washington, DC: The Urban Institute.

Harris, D., & Anderson, A. (2012). *Bias of public sector worker performance monitoring: Theory and empirical evidence from middle school teachers.* Paper presented at the 2012 annual meeting of the Association for Education Finance and Policy.

Harris, D., & Rutledge, S. (2010). Models and predictors of teacher effectiveness: A review of the evidence with lessons from (and for) other occupations. *Teachers College Record, 112,* 914–960.

Harris, D., Rutledge, S., Ingle, W., & Thompson, C. (2010). Mix and match: What principals really look for when hiring teachers. *Education Finance and Policy, 5,* 228–246.

Harris, D., & Sass, T. (2006). *Value-added models and the measurement of teacher productivity.* Paper presented at the annual meeting of the American Education Finance Association.

Harris, D., & Sass, T. (2009a). The effects of NBPTS-certified teachers on student achievement. *Journal of Policy Analysis and Management, 28,* 55–80.

Harris, D., & Sass, T. (2009b). *What makes for a good teacher and who can tell?* (CALDER Working Paper No. 30). Washington, DC: Urban Institute.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95,* 798–812.

Ingle, W. K., Rutledge, S. A., & Bishop, J. L. (2011). Context matters: Principals' sensemaking of teacher hiring and on-the-job performance. *Journal of Educational Administration, 49,* 579–610.

Jackson, C. K. (in press). Teacher quality at the high-school level: The importance of accounting for tracks. *Journal of Labor Economics.*

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources, 39,* 50–79.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26,* 101–136.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 187–200.

Kane, T. J., McCaffrey, D. M., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T., & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research, Inc.

Kennedy, M. (2004). Reform ideals and teachers' practical intentions. *Education Policy Analysis Archives, 12.* Retrieved from http://epaa.asu.edu/ojs/article/view/168/294

Kennedy, M. M. (2008). Sorting out teacher quality. *Phi Delta Kappan, 90*(1), 59–63.

Kennedy, M. (2010). Introduction: The uncertain relationship between teacher assessment and teacher quality. In M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook* (pp. 9–42). San Francisco, CA: Jossey-Bass.

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

Little, O. (2009). *Teacher evaluation systems: The window for opportunity and reform.* Washington, DC: National Education Association.

Liu, E., & Johnson, S. M. (2006). New teachers' experiences of hiring: Late, rushed, and information-poor. *Educational Administration Quarterly, 42*, 324–360.

Louis, K. S., Leithwood, K., Wahlstrom, K., & Anderson, S. (2010). *Investigating the links to improved student learning: Final report of research findings.* St. Paul, MN: University of Minnesota Center for Applied Research and Educational Improvement.

Marzano, R., Waters, T., & McNulty, B. (2011). *Leadership that works: From research to results.* Alexandria, VA: ASCD.

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

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research, 80*, 242–247.

Metzger, S. A., & Wu, M. J. (2008). Commercial teacher selection instruments: The validity of selecting teachers through beliefs, attitudes, and values. *Review of Educational Research, 78*, 921–940.

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

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

Monk, D. H. (1994). Subject matter preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*, 125–145.

Murnane, R. J. (1975). *The impact of school resources on the learning of inner-city children.* Cambridge, MA: Ballinger.

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis, 23*, 297–321.

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M. T., Filer, J. D., Wiedmaier, C. D., & Moore, C. W. (2007). Students' perceptions of characteristics of effective teachers: A validity study of a teaching evaluation form using mixed methods analysis. *American Educational Research Journal, 44*, 113–160.

Painter, S. R. (2000). Principals' efficacy beliefs about teacher evaluation. *Journal of Educational Administration, 38*, 368–378.

Parsons, T. (1960). *Structure and process in modern societies.* Glencoe, IL: Free Press.

Peterson, K. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal, 24*, 311–317.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices.* Thousand Oaks, CA: Corwin Press.

Ralph, E. G., Kesten, C., Lang, H., & Smith, D. (1998). Hiring new teachers: What do school districts look for? *Journal of Teacher Education, 49*, 1–10.

Rice, J. K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes.* Washington, DC: Economic Policy Institute.

Rivkin, S., Hanushek, E. A., & Kain, J. F. (2005). Teacher schools, and academic achievement. *Econometrica, 73,* 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94,* 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* (NBER Working Paper No. 16240). Cambridge, MA: National Bureau of Economic Research.

Rothstein, J. (2009). *Student sorting and bias in value added estimation: Selection on observables and unobservables* (NBER Working Paper No. 14666). Cambridge, MA: National Bureau of Economic Research, Inc.

Rutledge, S. A., Harris, D. N., Thompson, C. T., & Ingle, W. K. (2008). Certify, blink, hire: An examination of the process and tools of teacher screening and selection. *Leadership and Policy in Schools, 7,* 237–263.

Rutledge, S. A., Harris, D. N., & Ingle, W. K. (2010). How principals "bridge and buffer" the new demands of teacher quality and accountability. A mixed-methods analysis of teacher hiring. *American Journal of Education, 116,* 211–242.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57,* 1–22.

Snyder, T. D., Dillow, S. A., & Hoffman, C. M. (2007). *Digest of Education Statistics 2006 (NCES 2007-017).* Washington, DC: U.S. Government Printing Office.

Stodolosky, S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher, 13*(9), 11–18.

Strauss, R. P., & Vogt, W. B. (2006). *Should teachers know, or know how to teach?* Revised version of paper presented at the 2001 Annual Meeting of the American Educational Finance Association.

Theel, R. K., & Tallerico, M. (2004). Using portfolios for teacher hiring: Insights from principals. *Action in Teacher Education, 26*(1), 26–33.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal, 113,* F3–F33.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73,* 89–122.

Weisberg, D., Sexton, S., Mulhern, J., & Kelling, D. (2009). *The widget effect.* Retrieved from http://files.eric.ed.gov/fulltext/ED515656.pdf.

Wilson, S., & Floden, R. (2003). *Creating effective teachers: Concise answers for hard questions.* New York, NY: AACTE Publications.

Wilson, S., Floden, R., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations.* Seattle, WA: Center for the Study of Teaching and Policy, University of Washington.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432–442.

Manuscript received January 19, 2012
Final revision received July 15, 2013
Accepted August 9, 2013