CALDER

NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

Urban Institute

A program of research by the Urban Institute with Duke University, Stanford University, University of Florida,
University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

What Makes for a Good Teacher and Who Can Tell?

DOUGLAS N. HARRIS
AND TIM R. SASS

WORKING PAPER 30 • SEPTEMBER 2009

What Makes for a Good Teacher and Who Can Tell?

Douglas N. Harris
University of Wisconsin - Madison

Tim R. Sass
Florida State University

Corresponding authors: Douglas Harris, University of Wisconsin – Madison, *Email*: dnharris3@wisc.edu and Tim Sass, Florida State University*, Email:* tsass@fsu.edu.

The research presented here is funded under grant R305M040121 from the U.S. Department of Education. The authors wish to thank Stacey Rutledge, William Ingle, Peter Hinrichs, and participants in the NBER summer 2007 education workshop for their comments. John Gibson, William Ingle, Micah Sanders, Cynthia Thompson, and Lisa Verdon provided valuable research assistance. The authors are grateful for support from the National Center for Analysis of Longitudinal Data in Education Research (CALDER), supported by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education.

CALDER working papers have not gone through final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication.

The Urban Institute is a nonprofit, nonpartisan policy research and educational organization that examines the social, economic, and governance problems facing the nation. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Any errors are attributable to the authors.

CALDER, The Urban Institute
2100 M Street N.W., Washington, D.C. 20037
202-261-5739 ♦ www.caldercenter.org

Contents

What Makes for a Good Teacher and Who Can Tell?
Douglas N. Harris and Tim R. Sass
CALDER Working Paper No. 30
September 2009

Abstract

Mounting pressure in the policy arena to improve teacher productivity either by improving signals that predict teacher performance or through creating incentive contracts based on performance—has spurred two related questions: Are there important determinants of teacher productivity that are not captured by teacher credentials but that can be measured by subjective assessments? And would evaluating teachers based on a combination of subjective assessments and student outcomes more accurately gauge teacher performance than student test scores alone? Using data from a midsize Florida school district, this paper explores both questions by calculating teachers' "value added" and comparing those outcomes with subjective ratings of teachers by school principals. Teacher value-added and principals' subjective ratings are positively correlated and principals' evaluations are better predictors of a teacher's value added than traditional approaches to teacher compensation focused on experience and formal education. In settings where schools are judged on student test scores, teachers' ability to raise those scores is important to principals, as reflected in their subjective teacher ratings. Also, teachers' subject knowledge, teaching skill, and intelligence are most closely associated with both the overall subjective teacher ratings and the teacher value added. Finally, while past teacher value added predicts future teacher value added the principals' subjective ratings can provide additional information and substantially increase predictive power.

What Makes for a Good Teacher and Who Can Tell?

Introduction

Recent research consistently finds that teacher productivity is the most important component of a school's effect on student learning and that there is considerable heterogeneity in teacher productivity within and across schools (Rockoff 2004; Hanushek et al. 2005; Rivkin, Hanushek, and Kain 2005; Kane, Rockoff, and Staiger 2006; Aaronson, Barrow, and Sander 2007). Relatively little is known, however, about what makes some teachers more productive than others in promoting student achievement.

Older cross-sectional studies of educational production functions found that the characteristics that form the basis for teacher compensation—graduate degrees and experience—are at best weak predictors of a teacher's contribution to student achievement (Hanushek 1986, 1997). More recent estimates using panel data have determined that teacher productivity increases over the first few years of experience (Rockoff 2004; Clotfelter, Ladd, and Vigdor 2006; Jepsen 2005; Rivkin, Hanushek, and Kain 2005; Harris and Sass 2008; Aaronson, Barrow, and Sander 2007), but little else in the way of observed teacher characteristics seems to consistently matter.[1] In short, while teachers significantly influence student achievement, the variation in teacher productivity is still largely unexplained by commonly measured characteristics.

One possible explanation for the inability of extant research to identify the determinants of teacher productivity is that researchers have not been measuring the characteristics that truly

[1] Clotfelter, Ladd, and Vigdor (2007a, 2007b), using North Carolina data, find some teacher credentials are correlated with teacher effectiveness, particularly at the secondary level. Goldhaber (2007) also uses the North Carolina data and finds similar results, although he questions the signal value of credentials that are weakly correlated with productivity.

determine productivity. Previous studies in the economics of education literature have focused primarily on readily observed characteristics like experience, educational attainment, certification status, and college major. According to recent work in labor economics, however, personality traits may also play an important role in determining labor productivity (Borghans, ter Weel, and Weinberg 2008; Cunha et al. 2006; Heckman, Stixrud, and Urzua 2006).

Unraveling the factors that determine teacher productivity could yield valuable insights into the most appropriate policies for selecting and training teachers. If teacher productivity is affected primarily by personality characteristics that are measurable ex ante, they could be used as signals to identify the most desired candidates in the hiring process. If, however, valuable teacher characteristics are malleable, determining which teacher characteristics have the greatest impact on student learning could also inform the design of preservice and in-service teacher training programs.

Understanding the factors that affect teacher productivity and the degree to which these determinants are measurable would also inform current policy debates over how best to evaluate and compensate teachers. If it is not possible to measure the characteristics of teachers that determine their productivity, then ex post evaluation of teachers based on their contributions to student achievement or "value added" may be optimal (Gordon, Kane, and Staiger 2006). Currently, many school districts are experimenting with such "pay for performance" systems (Podgursky and Springer 2007), although some are concerned about the precision of these measures, their narrow focus on student test scores, and the fact that they can be calculated for only a small proportion of teachers. Alternatively, if there are particular teacher characteristics and behaviors that influence their productivity and are observable, but perhaps not easily quantified, then reliance on supervisor evaluations and other more subjective assessments may be advantageous. There is movement toward granting principals greater authority in hiring, evaluation, and retention of teachers both through the creation of independent charter schools nationwide and through decentralization reforms in public school districts such as New York City. The downside of subjective evaluations by principals is they may be affected by personal bias toward factors unrelated to productivity and some principals may simply be poor judges of teacher productivity.

To address the related issues of the determinants of teacher productivity and how best to evaluate teacher performance, we analyze the relationship between principal evaluations of teachers and the contribution of teachers to student achievement or teacher value added. Like other recent work, we examine the relationship between teacher characteristics (including typical teacher credentials and personality traits observed by principals) and both subjective and value-added measures of contemporaneous teacher performance. We also go beyond the existing research, however, and compare the ability of past value-added measures and principal ratings to predict *future* teacher value added.

We begin by estimating a model of student achievement that includes fixed effects to control for unmeasured student, teacher, and school heterogeneity. The resulting estimated teacher fixed effects are our measure of teacher value added. We then analyze the simple correlation between principals' subjective assessments and teachers' value-added scores. We follow that process with a multivariate analysis to examine whether principals are better at judging teacher productivity than traditional approaches to compensation that focus on experience and formal education. Next, we look in detail at specific teacher attributes and how they relate to both principals' overall evaluations of their faculty members and estimates of teacher value added. Finally, we compare the ability of principal ratings and past teacher value-added measures to predict future teacher value added.

In the next section, we describe the small literature on principal evaluations of teachers and their relationship with value added and follow with a discussion of the data used for our new analysis, including how we conducted the interviews with principals and our method for estimating teacher value added. In the concluding section, we discuss our empirical results and possible policy implications.

Literature Review

The literature of labor economics increasingly integrates theories and research from psychology. For example, Cuhna et al. (2006) model the life cycle of skill attainment, giving a prominent position to personality traits. 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 finding is perhaps not surprising, especially for jobs (such as teaching) that require substantial interpersonal interaction and communication, but it does suggest that economists may need to consider more than intelligence when evaluating the role of innate ability in labor market outcomes (Borghans, Duckworth, Heckman, and Bas ter Weel 2008).

The role of personality traits is also related to the way in which overall worker productivity is measured and rewarded in the workplace—in particular, the balance of subjective supervisor ratings and more objective measures of output. Research on the relationships between subjective and objective measures of worker productivity, as well as the implications of this relationship for optimal employment contracts, has a long history. As noted by Jacob and Lefgren (2008), this research suggests a relatively weak relationship between subjective and objective measures (Bommer et al. 1995; Heneman 1986). One reason might be that supervisors are heavily influenced by personality traits, more so than is warranted by the role personality actually plays in (objective) productivity. Evidence that evaluators' subjective assessments are biased—in the sense that certain types of workers (e.g., females and older workers) receive lower subjective evaluations for reasons that appear unrelated to their actual productivity (e.g., Varma and Stroh 2001)—reinforces this interpretation.

A limited literature specifically addresses the relationship between subjective and objective assessments of school teachers. Three older studies have examined the relationship between student test scores and principals' subjective assessments using longitudinal student achievement data to measure student learning growth (Murnane 1975; Armor et al. 1976; and Medley and Coker 1987). The use of panel data provides the opportunity to isolate teacher productivity from other time-invariant factors such as the unmeasured differences in student and family characteristics. As noted by Jacob and Lefgren (2008), however, these studies do not account for measurement error in the objective test–based measure and therefore understate the relationship between subjective and objective measures.

Jacob and Lefgren (2008) address both the selection bias and the measurement error problems within the context of a "value-added" model for measuring teacher productivity that is linked to principals' subjective assessments. They obtain student achievement data and combine it with data on principals' ratings of 201 teachers in a midsize school district in a Western state.[2] They find that principals can generally identify teachers who contribute the most and the least to student achievement but are less able to distinguish teachers in the middle of the productivity

[2] As in the present study, the district studied by Jacob and Lefgren (2008) chose to remain anonymous.

distribution.

Jacob and Lefgren (2008) also find that previous value added is a better predictor of current student outcomes than are current principal ratings. In particular, teacher value added calculated from test scores in 1998–2002 was a significantly better predictor of 2003 test scores (conditional on student and peer characteristics) than were 2003 principal ratings made just before the 2003 student exam. The principal ratings were also significant predictors of current test scores, conditional on prior value added. While this latter finding suggests that contemporaneous principal ratings add information, the reason is not clear. The principal ratings might provide more stable indicators of previous teacher productivity, since past value added is subject to transient shocks to student test scores. Alternatively, the principal ratings may simply reflect new current-school-year (2002–03) performance information not included in past value added (based on test scores through 2001–02). To sort out these effects, in our analysis we compare the ability of current value added and current principal ratings to predict future teacher value added.

The only prior study that considered principals' assessments of specific teacher characteristics, as opposed to the overall rating, is a working paper by Jacob and Lefgren (2005). Their list of items rated by principals includes both teacher characteristics and inputs (dedication and work ethic, organization, classroom management, providing a role model for students, positive relationships with teacher colleagues and administrators) and outputs (raising student achievement, student and parent satisfaction). They also apply factor analysis to these variables and create three broader variables: student satisfaction, achievement, and collegiality. The teacher's relationship with the school administration, however, is the only teacher characteristic they consider as a possible predictor of value added. (Their evidence suggests a positive and significant relationship between the two.)

A number of other studies have examined the relationship between the achievement levels of teachers' students and the subjective teacher ratings based on formal standards and extensive classroom observation (Gallagher 2004; Kimball et al. 2004; Milanowski 2004).[3] All these studies find a positive and significant relationship, despite differences in the way they measure teacher value added and in the degree to which the observations are used for high-stakes personnel decisions. While these studies have the advantage of more structured subjective evaluations, the reliance on achievement levels with no controls for lagged achievement or prior educational inputs makes it difficult to estimate teacher value added.

This study differs from Jacob and Lefgren (2008) and the previous literature as it is the first to study how well past subjective and objective ratings predict future productivity. We also build on previous work in other ways. First, we consider a broader range of teacher characteristics, one that is based on previous theories and evidence of teacher productivity.[4] We include personality traits such as "caring," "enthusiastic," and "intelligent," as well as evaluations of subject matter knowledge and teaching skill. Second, we analyze the relationship between each of these measures and both the overall evaluation by principals and the teacher value added. Finally, we analyze teacher ratings and student performance in middle and high school, in addition to elementary school, and allow the relationship between teacher characteristics and teacher ratings or teacher value added to vary across these grade groupings.

[3] For example, in Milanowski (2004), the subjective evaluations are based on an extensive standards framework that required principals and assistant principals to observe each teacher six times in total and, in each case, to rate the teacher on 22 separate dimensions.

[4] For a review of this literature, see Harris, Rutledge, Ingle, and Thompson (forthcoming)

Data and Methods

We begin by describing the general characteristics of the school district and sample of principals, teachers, and students. We then discuss in more detail the two main components of the data: (1) administrative data that are used to estimate teacher value added; and (2) principal interview data that provide information about principals' overall assessments of teachers as well as ratings of specific teacher characteristics.

General Sample Description

The analysis is based on interviews with 30 principals from an anonymous midsize Florida school district. The district includes a diverse population of students. For example, among the sampled schools, the average proportion of students eligible for free or reduced-price lunches varies from less than 10 percent to more than 90 percent. Similarly, there is considerable diversity among schools in the racial and ethnic distribution of their students. We interviewed principals from 17 elementary (or K–8) schools, 6 middle schools, 4 high schools, and 3 special-population schools, representing more than half the principals in the district. The racial distribution of interviewed principals is comparable to the national average of all principals (sample district, 78 percent white; national, 82 percent white) as is the percentage with at least a master's degree (sample district, 100 percent; national, 90.7 percent).[5] The percentage female, however, is somewhat larger (sample district, 63 percent; national, 44 percent).

The advantage of studying a school district in Florida is that the state has a long tradition of strong test-based accountability (Harris, Herrington, and Albee 2007) that is now coming to

[5] The national data on principals comes from the 2003–2004 Schools and Staffing Survey (SASS) as reported in the Digest of Education Statistics (National Center for Education Statistics 2006). 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.

pass in other states as a result of No Child Left Behind. The state has long graded schools on an A–F scale. The number of schools receiving the highest grade has risen over time; in our sample, 20 schools received the highest grade (A) during the 2005–06 school year, and the lowest grade was a D (one school). It is reasonable to expect that accountability policies, such as the school grades mentioned above, influence the objectives that principals see for their schools and therefore their subjective evaluations of teachers. For example, we might expect a closer relationship between value added and subjective assessments in high accountability contexts not only where principals are more aware of test scores in general but also where they are increasingly likely to know the test scores, and perhaps test score gains, made by students of individual teachers. We discuss the potential influence of this phenomenon later in the analysis but emphasize here that, by studying a Florida school district, the results of our analysis are more applicable to the current policy environment where high-stakes achievement-focused accountability is federal policy.

Student Achievement Data and Modeling

Florida conducts annual testing in grades 3 through 10 for both math and reading. Two tests are administered, a criterion-referenced exam based on the state curriculum standards known as the FCAT-Sunshine State Standards exam, and the norm-referenced Stanford Achievement Test. We employ the Stanford Achievement Test in the present analysis for two reasons. First, 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. Second, and most important, the district under study also administers the Stanford Achievement Test in grades 1 and 2, allowing us to compute achievement gains for students in grades 2 through 10. Achievement data on the Stanford

Achievement Test are available for each of the school years 1999–2000 through 2007–08.[6] Thus, we are able to estimate the determinants of achievement gains for five years before the principal interviews, 2000–01through 2005–06, and for two years after the interviews, 2006–07 through 2007–08. Characteristics of the sample used in the value-added analysis are described in table 1.

Table 1. Sample Student and Teacher Characteristics

	Math Sample		Reading Sample	
	Observations (no.)	Mean	Observations (no.)	Mean
Students				
Black	31,645	0.367	30,794	0.360
Hispanic	31,645	0.025	30,794	0.024
Free or reduced-price lunch	31,645	0.335	30,794	0.329
Achievement gain	31,645	20.729	30,794	18.581
Teachers				
Male	1,023	0.115	1,024	0.079
White	1,023	0.695	1,024	0.724
Hold advanced degree	1,004	0.332	1,008	0.350
Fully certified	1,015	0.950	1,019	0.955
Taught primarily elementary school	1,023	0.727	1,024	0.729
Taught primarily middle school	1,023	0.149	1,024	0.141
Taught primarily high school	1,023	0.124	1,024	0.130
Principal's overall rating	237	7.084	231	7.134
Rating of ability to raise test scores	210	7.200	201	7.184
Rating on "caring"	237	7.384	231	7.463
Rating on "enthusiastic"	237	7.249	231	7.372
Rating on "motivated"	237	7.414	231	7.481
Rating on "strong teaching skills"	237	7.544	231	7.636
Rating on "knows subject"	237	7.848	231	7.918
Rating on "communication skills"	237	7.612	231	7.758
Rating on "intelligence"	237	7.911	231	7.970
Rating on "positive relationship with parents"	236	7.483	230	7.600
Rating on "positive relationship with students"	236	7.636	230	7.739

Note: Includes only students and teachers for which a fixed effect could be computed for the teacher.

[6] Before 2004–2005, version 9 of the Stanford Achievement Test (SAT-9) was administered. Beginning in 2004–2005, the SAT-10 was given. All SAT-10 scores have been converted to SAT-9 equivalent scores based on the conversion tables in Harcourt (2002).

To compute value-added scores for teachers, we estimate a model of student achievement of the following form:

$$\Delta A_{it} = \boldsymbol{\beta}_1 \mathbf{X}_{it} + \boldsymbol{\beta}_2 \mathbf{P}_{-ijmt} + \gamma_i + \delta_k + \phi_m + \nu_{it} \qquad (1)$$

The vector $\boldsymbol{X}_{it}$ includes time-varying student characteristics such as student mobility. The vector of peer characteristics, $\boldsymbol{P}_{-ijmt}$ (where the subscript *–i* students other than individual *i* in the classroom), includes both exogenous peer characteristics and the number of peers or class size. There are three fixed effects in the model: a student fixed effect (γ_i), a teacher fixed effect (δ_k), and a school fixed effect, ϕ_m. The teacher fixed effect captures the time-invariant characteristics of teachers as well as the average value of time-varying characteristics like experience and possession of an advanced degree. Since school fixed effects 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, ν_{it}, is a mean zero random error. The model is based on the cumulative achievement model of Todd and Wolpin (2003) and derived in detail in Harris and Sass (2006).

Recently, Rothstein (2009) has argued that such value-added models may produce biased estimates of teacher productivity because of the nonrandom assignment of students to teachers within schools. While our use of student fixed effects controls for sorting based on time-invariant characteristics, Rothstein argues that teacher effect estimates could still be biased if teacher assignments are determined by transient shocks to student achievement. For example, if students who experience an unusually high achievement gain in one year are assigned to particular teachers the following year and there is mean reversion in student test scores, the estimated value added for the teachers with high prior-year gains will be biased downward. Rothstein proposes falsification tests based on the idea that future teachers cannot have causal effects on current achievement gains. We conduct falsification tests of this sort, using the methodology employed by Koedel and Betts (2009). For each level of schooling—elementary, middle, and high—we fail to reject the null of strict exogeneity, indicating that the data from the district we analyze in this study are not subject to the sort of dynamic sorting bias concerns raised by Rothstein.[7]

As noted by Jacob and Lefgren (2008), another concern is measurement error in the estimated teacher effects. Given the variability in student test scores, value-added estimates will yield "noisy" measures of teacher productivity, particularly for teachers with relatively few students (McCaffrey et al. forthcoming). We employ two strategies to alleviate the measurement error problem. First, we limit our sample to teachers who taught at least five students with achievement gain data. Second, we compute empirical Bayes "shrunken" estimates of teacher productivity, which are essentially a weighted average of the individual teacher-effect estimates and the average teacher-effect estimate, with greater weight given to the individual estimates the smaller is their standard error.[8] As noted by Mihaly et al. (2009), standard fixed-effects software routines compute fixed effects relative to some arbitrary hold-out unit (e.g., an omitted teacher), which can produce wildly incorrect standard errors and thus inappropriate empirical Bayes estimates. Therefore, to estimate the teacher effects and their standard errors, we employ the Stata routine *felsdvregdm*, developed by Mihaly et al. (2009), which imposes a sum-to-zero constraint on the estimated teacher effects within a school and produces the appropriate standard errors for computing the empirical Bayes (shrunken) estimates of teacher value added.

[7] For math, the *p*-values on the test of zero future teacher "effects" were 1.00 for elementary school, 0.75 for middle school, and 0.63 for high school. For reading, the corresponding *p*-values were 1.00, 0.35, and 0.20.

[8] For details on the computation of empirical Bayes estimates, see Morris (1983) and Jacob and Lefgren (2005).

Principal Interview Data

We conducted interviews with the principals in the summer of 2006, asking each principal to rate up to 10 teachers in grades and subjects that are subject to annual student achievement testing. Per the requirements of the district, the interviews were "single-blind" so that the principal knew the names of the teachers but the interviewer knew only a randomly assigned number associated with the names.

From the administrative data described above, we identified teachers in tested grades and subjects in the 30 schools who had taught at least one course with 10 or more tested students and who were still in the school in the 2004–05 school year (the last year for which complete administrative data were available before we conducted the principal interviews). In some cases, there were fewer than 10 teachers who met these requirements. Even in schools that had 10 teachers on the list, some teachers were not actually working in the respective schools at the time of the interview. If the principal was familiar with a departed teacher and felt comfortable making an assessment, then these teachers and subjective assessments were included in the analysis. If the principal was not sufficiently familiar with the departed teacher, then the teacher was dropped. Many schools had more than 10 teachers. 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.

The first question in the interview asked the principals to mark on a sheet of paper his or her overall assessment of each teacher, using a 1-to-9 scale.[9] The interviewer then handed the principal another sheet of paper so that he or she could rate each teacher on each of 12

[9] The specific question was, "First, I would like you to rate each of the ten teachers relative to the other teachers on the list. Please rate each teacher on a scale from 1–9 with 1 being not effective to 9 being exceptional. Place an X in the box to indicate your choice. Also please circle the number of any teachers whose students are primarily special populations."

characteristics: caring, communication skills, enthusiasm, intelligence, knowledge of subject, strong teaching skills, motivation, works well with grade team and 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 by Harris, Rutledge, Ingle, and Thompson (forthcoming) to be among the most important characteristics that principals look for when hiring teachers.[10] Having an occupation-specific list of characteristics is important because recent economic theory and evidence suggest that different traits matter more in different occupations (Borghans, ter Weel, and Weinberg 2008) and specifically that "caring" is more important in teaching than in any other occupation.

The interview questions were designed so that principals would evaluate teachers relative to others in the school, since even an "absolute" evaluation would be necessarily based on each principal's own experiences. Ratings on individual characteristics across principals, therefore, may not be based on a common reference point or a common scale. In our analyses, then, like Jacob and Lefgren (2008), we normalize the ratings of each teacher characteristic to have a mean of zero and standard deviation of one over all teachers rated by a given principal. Given that our teacher fixed-effects estimates are within-school measures, normalizing the ratings allows us to compare within-school ratings to within-school teacher value added.

[10] As described in Harris, Rutledge, Ingle, and Thompson (forthcoming), the data in this study came from the second in a series carried out by the researchers. During the summer of 2005, interviews were conducted on the hiring process and on the principals' preferred characteristics of teachers. The first set of interviews is important because it helps validate the types of teacher characteristics we consider. Principals were asked an open-ended question about the teacher characteristics they prefer. Two-thirds of these responses could be placed in one of 12 categories identified from previous studies on teacher quality. The list here takes those ranked highest by principals in the first interview and then adds some of those included by Jacob and Lefgren (2008).

In the final activity of the interview, the principals rated each teacher according to the following additional "outcome" measures: "raises FCAT math achievement," "raises FCAT reading achievement," "raises FCAT writing achievement," "positive relationship with parents," and "positive relationship with students." These last measures help us test whether the differences between the value-added measures and the principals' overall assessments are due to philosophical differences on the importance of student achievement as an educational outcome or to difficulty in identifying teachers who increase student test scores.

Finally, as part of the interview, we discovered that principals have access to a district-purchased software program, Snapshot, that allows them to create various cross-tabulations of student achievement data, including simple student learning gains and mean learning gains by teacher. While we have no data about the actual use of this software, subsequent informal conversations with two principals suggest that at least some principals use the program to look at the achievement gains made by students of each teacher. While this resource may have provided principals with some information about unconditional student average achievement gains, that is. of course, not the same thing as the teacher value-added scores, which are conditional on student and peer characteristics.

Results

To compute value-added scores for teachers, we estimate equation (1) using data on test score gains for grades 2–10 over the period 2000–01 through 2005–06. To lessen potential multicollinearity problems and reduce the number of teacher characteristics to analyze, we follow Jacob and Lefgren (2005) and conduct a factor analysis of the 11 individual teacher characteristics rated by principals. As indicated in table 2, the individual characteristics can be

Table 2. Factor Loadings of Normalized Principal Ratings

Teacher characteristic rated by principal	Interpersonal skills	Motivation, enthusiasm	Works well with others	Knowledge, teaching skills, intelligence
Math				
Intelligent	-0.0481	0.0839	0.0606	**0.7067**
Works well with grade team/dept.	-0.0046	-0.0887	**0.9711**	0.0399
Works well with me (principal)	0.1743	0.0835	**0.7415**	-0.0814
Positive relationship with parents	**0.7231**	0.0781	0.0768	0.0742
Positive relationship with students	**0.9408**	0.0103	-0.0131	0.0636
Caring	**0.5591**	0.1372	0.2422	-0.0185
Enthusiastic	0.1086	**0.9721**	-0.0707	-0.0035
Motivated	0.0398	**0.5224**	0.2802	0.1624
Strong teaching skills	0.1512	0.0258	-0.0462	**0.8471**
Knows subject	-0.0088	-0.0551	-0.0036	**0.9831**
Communication skills	0.1040	0.1705	0.2734	0.3191
Reading				
Intelligent	-0.0138	0.0094	0.0445	**0.7064**
Works well with grade team/dept.	0.0179	-0.0581	**0.8646**	0.0704
Works well with me (principal)	0.1507	0.0409	**0.8251**	-0.0558
Positive relationship with parents	**0.7559**	0.0511	0.0637	0.0741
Positive relationship with students	**0.9195**	0.0258	0.0181	0.0287
Caring	**0.5970**	0.0989	0.2610	-0.0385
Enthusiastic	0.0728	**0.9942**	-0.0476	-0.0225
Motivated	0.0728	**0.5289**	0.1894	0.2529
Strong teaching skills	0.2269	0.0127	-0.0854	**0.8175**
Knows subject	-0.0814	-0.0201	0.0333	**0.9840**
Communication skills	0.1484	0.2225	0.1855	0.3214

Notes: Principal ratings are normalized within principal to have mean zero and variance of one. Factor analysis uses maximum likelihood method. Factor loadings are based on promax rotation. Numbers in bold indicate most important components of each factor.

summarized into four factors: interpersonal skills, motivation and enthusiasm, ability to work with others, and knowledge, teaching skills, and intelligence.

Simple correlations among the estimated teacher fixed effects, principals' overall ratings of teachers, principals' ratings of a teacher's ability to raise test scores on the relevant achievement test, and the four teacher characteristic factors are presented in table 3. The first column shows positive relationships between teacher value added and all the teacher characteristic factors. The overall principal rating is positively associated with value added,

Table 3. Pairwise Correlation of Estimated Teacher Fixed Effects and Principal's Rating of Teachers with Teacher Characteristic Factors

	Estimated teacher fixed effect	Overall rating	Ability to raise test scores	Interpersonal skills	Motivation, enthusiasm	Works well with others	Knowledge, teaching skills, intelligence
Math							
Estimated teacher fixed effect	1.000						
Overall rating	0.276 **	1.000					
Ability to raise test scores	0.265 **	0.733 **	1.000				
Interpersonal skills	0.212 **	0.703 **	0.550 **	1.000			
Motivation, enthusiasm	0.188 **	0.738 **	0.596 **	0.734 **	1.000		
Works well with others	0.204 **	0.762 **	0.598 **	0.756 **	0.732 **	1.000	
Knowledge, teaching skills, intelligence	0.269 **	0.881 **	0.752 **	0.612 **	0.682 **	0.644 **	1.000
Reading							
Estimated teacher fixed effect	1.000						
Overall rating	0.219 **	1.000					
Ability to raise test scores	0.214 **	0.741 **	1.000				
Interpersonal skills	0.143 **	0.709 **	0.626 **	1.000			
Motivation, enthusiasm	0.163 **	0.697 **	0.569 **	0.716 **	1.000		
Works well with others	0.099	0.723 **	0.589 **	0.763 **	0.676 **	1.000	
Knowledge, teaching skills, intelligence	0.209 **	0.856 **	0.702 **	0.632 **	0.684 **	0.650 **	1.000

** indicates significance at the .05 level.

though, as in previous studies, this relationship is relatively weak. The correlation between value added and the principal's impression of a teacher's ability to raise test scores (the subjective equivalent of value added) is similarly low. One possible explanation is that principals evaluate a teacher based on simple mean gains in student test scores, rather than value added, which represents gains conditional on student and peer characteristics.

The relatively high correlation of 0.7 between principals' overall rating and their ratings on ability of teachers to raise test scores suggests that principals weigh the ability of teachers to boost student test scores highly in their overall evaluation. These findings hold for both math and reading. It is also noteworthy that the teacher-characteristics factors are all positively correlated with one another and are often highly correlated. It is not obvious that this should be the case—for example, that teachers who are more knowledgeable would also tend to have better interpersonal skills. It is possible there is a "halo effect" whereby teachers who are rated highly by the principal overall are automatically given high marks on all the individual characteristics, although this is very difficult to test without having some other independent measure of teacher characteristics. Finally, note that, among the four teacher characteristic factors, "knowledge, teaching skills, intelligence" is most closely associated with teacher value added. This result holds up in the regression analyses below.

Table 4 presents estimates of the determinants of the teacher fixed effects, including only standard teacher credentials (experience, possession of an advanced degree, certification status) along with general principal evaluations (overall rating, ability to raise test scores) as explanatory variables. The first column reports estimates where only teacher credentials and no principal ratings are included. With the exception of one of the experience measures, none of the

Table 4. Ordinary Least Squares Estimates of Determinants of Teacher Fixed Effects

	Math			Reading		
	[1]	[2]	[3]	[1]	[2]	[3]
Overall rating		2.374 ***			0.858 ***	
		[4.50]			[3.26]	
Ability to raise test scores			2.199 ***			0.845 ***
			[3.83]			[2.94]
1–2 years of experience	8.424	11.133	10.655	1.404	1.491	0.360
	[1.18]	[1.62]	[1.42]	[0.46]	[0.50]	[0.09]
3–5 years of experience	7.770	8.690	10.187 *	2.551	2.343	1.789
	[1.47]	[1.71]	[1.85]	[1.05]	[0.99]	[0.56]
6–12 years of experience	7.255	9.111 *	8.863 *	1.627	1.829	1.031
	[1.43]	[1.86]	[1.68]	[0.69]	[0.79]	[0.33]
13–20 years of experience	7.579	9.641 *	10.615 **	1.904	2.064	1.236
	[1.48]	[1.96]	[2.00]	[0.80]	[0.89]	[0.39]
21–27 years of experience	10.685 **	12.001 **	12.93 **	2.492	2.307	2.046
	[2.08]	[2.43]	[2.44]	[1.05]	[1.00]	[0.64]
28+ years of experience	8.539	10.567 **	10.184 *	2.606	2.739	1.734
	[1.63]	[2.10]	[1.87]	[1.08]	[1.16]	[0.54]
Advanced degree	-1.534	-1.532	-1.175	-0.053	-0.092	-0.399
	[1.38]	[1.44]	[1.04]	[0.10]	[0.17]	[0.09]
R-squared	0.039	0.117	0.126	0.016	0.06	0.061
Number of observations	237	237	202	231	231	201

Notes: Absolute values of t-ratios appear in brackets. * indicates statistical significance at .10 level, **indicates significance at the .05 level, and *** indicates significance at the .01 level in a two-tailed test. All models include a constant term.

credential variables is a statistically significant determinant of teacher value-added scores.[11] None of the coefficients is significant in the first column for reading, likely because of the relatively small sample size, as other statewide studies in Florida do show positive coefficients on experience (Harris and Sass 2008).

[11] In another study using statewide data from Florida (Harris and Sass 2008), the effects of teacher experience are highly significant when teacher fixed effects are excluded, but within-teacher changes in experience are less often statistically significant. The finding that experience is insignificant in models with teacher fixed effects could mean that some apparent cross-teacher experience effects are due to attrition of less effective teachers early in their careers or that there is simply insufficient within-teacher variation in experience over a short panel. The large estimated coefficients here for full certification of reading teachers are likely picking up idiosyncratic features of the handful of reading teachers in the sample who are not fully certified during part of the sample period.

In contrast, when a principal's overall rating of a teacher is added to the model, its coefficient is positive and highly significant in both reading and math. (The coefficients on teacher credentials are largely unchanged.) This finding suggests that principals have knowledge about teacher productivity that is not captured by the standard measures of experience, educational attainment, and certification that typically form the basis for teacher pay scales.

It is common to interpret the magnitude of coefficients in these types of models in terms of student-level standard deviations. For example, the coefficient on principals' overall ratings for math teachers in table 4 is +2.374, which implies that a teacher who is rated one point higher on the 1–9 scale raises student math test scores by 2.374 scale score points per year more than the average teacher, which translates to 0.04 student test score standard deviations.[12] While this might be considered small by some standards, these represent only single-year changes, which could accumulate to relatively larger effects over time.

In table 5 we present estimates where the correlation between principal ratings and estimated teacher value added is allowed to vary between elementary school and middle or high school. At the elementary level, the two principal ratings ("overall" and "ability to raise test scores") are positively and statistically significantly associated with the teacher fixed effect in both reading and in math. The effect of a one-point increase in the principal's rating scale on teacher value added in reading, however, is about half the size of the effect in math. This result is consistent with the general finding in the literature that the effects of teacher characteristics on student achievement tend to be less pronounced in reading. It is often suggested that reading

[12] This conversion is based on the standard deviation in the level of math achievement, 53.26. The standard deviation in the level of reading achievement is 50.58. Boyd et al. (2008) argue for measuring teacher effects relative to the standard deviation of student gains. This would roughly double the effect sizes, as the standard deviation of achievement gains are 23.20 for reading and 20.64 for math.

Table 5. Ordinary Least Squares Estimates of the Determinants of Teacher Fixed Effects

	Math			Reading		
	[1]	[2]	[3]	[1]	[2]	[3]
Overall rating x elementary		2.956 ***			1.072 ***	
		[4.91]			[3.53]	
Overall rating x middle/high		0.524			0.232	
		[0.49]			[0.45]	
Ability to raise test scores x elementary			2.967 **			1.21 **
			[4.51]			[3.52]
Ability to raise test scores x middle/high			0.059			0.0145
			[0.05]			[0.03]
1–2 years of experience	8.424	10.970	10.693	1.404	1.914	1.475
	[1.18]	[1.61]	[1.44]	[0.46]	[0.63]	0.36]
3–5 years of experience	7.770	8.141	9.947 *	2.551	2.471	2.331
	[1.47]	[1.61]	[1.83]	[1.05]	[1.04]	[0.73]
6–12 years of experience	7.255	9.103 *	8.647 *	1.627	2.067	1.845
	[1.43]	[1.87]	[0.65]	[0.69]	[0.90]	[0.58]
13–20 years of experience	7.579	9.115 *	10.045 *	1.904	2.222	1.933
	[1.48]	[1.86]	[1.91]	[0.80]	[0.96]	[0.61]
21–27 years of experience	10.685 **	4.911 **	12.454 **	2.492	2.599	2.659
	[2.08]	[2.39]	[2.38]	[1.05]	[1.12]	[0.84]
28+ years of experience	8.539	10.111 **	10.042 *	2.606	2.927	2.513
	[1.63]	[2.02]	[1.86]	[1.08]	[1.24]	[0.78]
Advanced degree	-1.534	-1.468 *	-1.199	-0.053	-0.041	-0.290
	[1.38]	[1.74]	[1.07]	[0.10]	[0.08]	[0.50]
R-squared	0.039	0.132	0.149	0.016	0.069	0.079
Number of observations	237	237	202	231	231	201

Notes: Absolute values of t-ratios appear in brackets. * indicates statistical significance at .10 level, ** indicates significance at the .05 level, and *** indicates significance at the .01 level in a two-tailed test. All models include a constant term.

scores are more likely to be influenced by factors outside of school; students may read books in their free time, but they seldom work math problems for enjoyment.

For middle and high school teachers, there are no significant relationships.[13] This difference may reflect difficulties in aligning the content of reading exams with the teacher responsible for the relevant instruction in higher grade levels. In elementary schools, the matching of courses to the content of reading exams is relatively easy because students typically have only one teacher. In middle and high school, however, literature courses may cover much

[13] While there are fewer middle and high school teachers than elementary teachers in the sample, the insignificant effects for middle and high school teachers are not due to the sample size. We restricted the sample so that the number of elementary teachers equaled the number of middle and high school teachers and obtained similar results.

material other than reading instruction, and reading scores may be influenced by classes such as social studies, which involve reading but where developing reading is not the primary purpose.[14]

We next turn to an analysis of the factors affecting a principal's overall rating of a teacher. Table 6 presents least-squares estimates from regressing the principal's overall rating on the perceived ability to raise test scores in the relevant subject and the principal's overall rating of teachers. For both math and reading, ability to raise test scores is highly correlated with the overall rating. This result is true for all teachers as well as for the subgroups of elementary and middle and high school teachers. There is more to the overall rating than ability to raise test scores, however; about 45 percent of the variation in overall ratings is due to other factors.

Table 6. OLS Estimates of the Determinants of Principal's Overall Rating of Teachers

	Math		Reading	
	[1]	[2]	[1]	[2]
Ability to raise test scores	0.733 ***		0.73 ***	
	[14.59]		[15.18]	
Ability to raise test scores x elementary		0.755 ***		0.771 ***
		[12.95]		[13.33]
Ability to raise test scores x middle/high		0.67 ***		0.637 ***
		[6.95]		[7.26]
R-squared	0.539	0.54	0.564	0.568
Number of observations	202	202	201	201

Notes: Absolute values of t-ratios appear in brackets. * indicates statistical significance at .10 level, ** indicates significance at the .05 level, and *** indicates significance at the .01 level in a two-tailed test. All models include controls for teacher experience, attainment of an advanced degree, and a constant term.

To determine what specific factors influence a principal's overall rating of a teacher, we reestimate the teacher rating model using the principal's rating of the four teacher characteristic factors. The results are presented in table 7. In both subjects, the "knowledge, teaching skills, and intelligence" criterion contributes the most to the principals' overall rating. While "works well with others" and "interpersonal skills" are statistically significant, the point estimates are much

[14] Koedel (2009) provides some evidence that social studies teachers influence reading test scores at the high school level.

Table 7. OLS Estimates of the Determinants of Principal's Overall Rating of Teachers, Allowing Effects to Vary Across Grade Groups

	Math		Reading	
	[1]	[2]	[1]	[2]
Interpersonal skill	0.096 **		0.187 ***	
	[2.05]		[3.32]	
Knowledge, teaching skills, intelligence	0.609 ***		0.601 ***	
	[15.36]		[12.57]	
Motivation, enthusiasm	0.054		0.024	
	[1.19]		[0.46]	
Works well with others	0.233 ***		0.156 ***	
	[4.94]		[2.88]	
Interpersonal skill x elementary		0.108 **		0.13 **
		[2.08]		[2.10]
Interpersonal skill x middle/high		0.051		0.505 ***
		[0.41]		[3.69]
Knowledge, teaching skills, intelligence x elementary		0.615 ***		0.613 ***
		[13.80]		[12.20]
Knowledge, teaching skills, intelligence x middle/high		0.599 ***		0.44 ***
		[6.26]		[3.03]
Motivation, enthusiasm x elementary		0.043		0.056
		[0.87]		[0.97]
Motivation, enthusiasm x middle/high		0.176		-0.048
		[1.30]		[0.44]
Works well with others x elementary		0.248 ***		0.19 ***
		[4.78]		[3.28]
Works well with others x middle/high		0.112		-0.031
		[0.89]		[0.22]
R-squared	0.852	0.854	0.805	0.814
Number of observations	207	207	203	203

Note: Absolute values of t-ratios appear in brackets. * indicates statistical significance at .10 level, **indicates significance at the .05 level, and *** indicates significance at the .01 level in a two-tailed test. All models include controls for teacher experience, attainment of an advanced degree, and a constant term.

smaller. There are some apparent differences by grade level, although none of these differences is statistically significant. Also, note that four factors explain roughly 80 percent of the variation in overall ratings, suggesting that the underlying 12 characteristics are important determinants of principals' overall assessments.

Very different patterns emerge when we switch the dependent variable to teacher value added in table 8. Column [1] suggests that "knowledge, teaching skills, intelligence" is positively and significantly associated with teacher value added in reading and math. None of the other coefficients in column [1] are significant. Column [2] shows that the effect of knowledge,

Table 8. OLS Estimates of the Determinants of Teacher Fixed Effects, Allowing Effects to Vary across Grade Groups

	Math		Reading	
	[1]	[2]	[1]	[2]
Interpersonal skill	0.751		0.276	
	[0.79]		[0.55]	
Knowledge, teaching skills, intelligence	1.792 **		0.783 *	
	[2.23]		[1.83]	
Motivation, enthusiasm	-0.105		0.280	
	[0.11]		[0.60]	
Works well with others	0.143		-0.534	
	[0.15]		[1.11]	
Interpersonal skill x elementary		0.807		0.231
		[0.77]		[0.41]
Interpersonal skill x middle/high		2.359		1.067
		[0.95]		[0.86]
Knowledge, teaching skills, intelligence x elementary		2.200 **		0.949 **
		[2.45]		[2.09]
Knowledge, teaching skills, intelligence x middle/high		-0.149		-0.207
		[0.08]		[0.16]
Motivation, enthusiasm x elementary		0.139		0.402
		[0.14]		[0.77]
Motivation, enthusiasm x middle/high		-1.958		-0.205
		[0.72]		[0.21]
Works well with others x elementary		0.215		-0.480
		[0.21]		[0.92]
Works well with others x middle/high		0.408		-0.794
		[0.16]		[0.63]
R-squared	0.128	0.149	0.069	0.085
Number of observations	207	207	203	203

Note: Absolute values of t-ratios appear in brackets. * indicates statistical significance at .10 level, **indicates significance at the .05 level, and *** indicates significance at the .01 level in a two-tailed test. All models include controls for teacher experience, attainment of an advanced degree, and a constant term.

teaching skills, and intelligence is entirely in the elementary grades. The overall explanatory power of the four factors is quite low, however.[15]

To this point, we have been using achievement data up through the 2005–06 school year to compute teacher value added, and we compared this value-added measure with various principal ratings of their teachers obtained in the summer of 2006. Such contemporaneous estimates are relevant to decisions about the role of principal evaluations in measuring and

[15] Some of the insignificant effects may be due to multicollinearity. As demonstrated in table 5, the four factors are all positively correlated. When each factor is regressed on estimated teacher effects separately, all are significant except "works well with others" in predicting the value-added of reading teachers.

rewarding past performance. Contemporaneous measures of teacher performance, however, are not particularly relevant for retention and tenure decisions, where the decision should (optimally) be based on predictions about future performance.

We measure future teacher productivity by re-estimating equation (1), using data on student achievement gains from the 2006–07 and 2007–08 school years (including test scores from 2005–06 as the initial lagged value), to derive estimates of future teacher value added. As demonstrated by (McCaffrey et al. forthcoming), basing teacher value added on two years of performance leads to much more precise estimates than relying on a single estimated test score gain, as in Jacob and Lefgren (2008). We then regress our estimate of future value added, which uses two estimated gains, on either the principal's overall rating of the teacher from the summer of 2006 or the estimated teacher fixed effect from a student achievement model covering the years 1999–2000 to 2005–06. As shown in table 9, we estimate the equation several ways, varying the amount of information used to estimate the past teacher value added.

When entered separately, past value added and past principals' ratings are positive and significant predictors of future teacher value added, no matter how past value added is estimated. The relative performance of the two measures in predicting future value added, however, varies directly with the amount of information used to compute prior value added. Using all available information, we find that past value added outperforms principal ratings, explaining eight times as much of the variation in future value added among math teachers and nearly twice as much of the variation among reading teachers. While the edge in explanatory power holds up in math when only two prior years of data are used to compute past value added, the difference is eliminated in reading. When past value added is based on a single year of data, principal ratings (which are typically based on multiple years of observation) outperform past value added in

Table 9. Estimates of the Determinants of Teacher Effects in 2006/07–2007/08

	Math			Reading		
	[1]	[2]	[3]	[1]	[2]	[3]
Prior value-added based on up to six years of teacher performance						
Prior value-added (2000/01–2005/06)	0.346 ***		0.336 ***	0.333 ***		0.284 ***
	[5.69]		[5.30]	[3.31]		[2.75]
Principal's overall rating (summer 2006)		1.219 **	0.362		1.309 **	0.946 *
		[1.97]	[0.61]		[2.59]	[1.85]
R-squared	0.162	0.023	0.163	0.070	0.044	0.091
Number of observations	170	170	170	149	149	149
Prior value-added based on up to two years of teacher performance						
Prior value-added (2004/05–2005/06)	0.309 ***		0.298 ***	0.482 **		0.399 *
	[5.52]		[5.28]	[2.35]		[1.94]
Principal's overall rating (summer 2006)		1.209 *	0.755		1.328 **	1.150 **
		[1.95]	[1.30]		[2.62]	[2.25]
R-squared	0.154	0.022	0.163	0.036	0.045	0.069
Number of observations	169	169	169	148	148	148
Prior value-added based on up to one year of teacher performance						
Prior value-added (2005/06)	-0.011		-0.01	6.428		1.602
	[1.04]		[0.93]	[0.64]		[0.16]
Principal's overall rating (summer 2006)		1.609 **	1.569 **		1.699 ***	1.680 **
		[2.16]	[2.10]		[2.70]	[2.61]
R-squared	0.008	0.033	0.039	0.004	0.058	0.058
No. of Observations	140	140	140	120	120	120

Notes: Data apply only to teachers teaching in same school in which they were previously rated by principal. Absolute values of t-ratios appear in brackets. * indicates statistical significance at .10 level, **indicates significance at the .05 level, and *** indicates significance at the .01 level in a two-tailed test. All models include a constant term.

predicting the future value added of teachers. This finding reinforces the importance of using multiple years of data to estimate teacher value added.

When prior value added and principal ratings are combined to predict future teacher performance, the principal ratings always add some information, although their contribution to the predictive power of the model depends on the precision of the past value-added measure. When past value added is based on all six years of achievement gain data before summer 2006, principal ratings add virtually nothing to the predictive power of past value added in math but increase the proportion of variation in the future value added of reading teachers from 7 percent to 9 percent. As fewer data are used to construct prior value-added estimates, the relative contribution of principal ratings grows. For example, when two years of data are used to compute prior value added, principal ratings increase the proportion of variation in future value added explained by about one percentage point in math and nearly double the proportion of variation explained in reading.

Summary and Conclusions

Consistent with prior research, we find that estimates of teachers' contributions to student achievement or "value added" are at best weakly correlated with readily observable teacher characteristics like experience and attainment of advanced degrees, suggesting that other factors may be relatively more important in determining what makes a "good" teacher. Teacher value added is correlated with traditional human capital measures like teacher intelligence, subject knowledge, and teaching skills, while personality traits like caring, motivation, enthusiasm, and ability to work well with others are not significantly related to teacher productivity in raising student achievement. In contrast, principal evaluations of teachers appear to be based on a broader set of characteristics, encompassing teacher knowledge, skill, and intelligence but also including interpersonal relationships with parents, other teachers, and the principal and a caring attitude toward students.

The divergence in the factors associated with teacher value added and those that are related to principal evaluation may arise in part because principals consider educational objectives beyond student achievement. We find differences between principals' overall assessments of teachers and principals' impressions of how well the same teachers raise student achievement. The correlation between the two assessments is relatively strong, however. This connection may reflect both the principals' desire to be consistent in their various ratings of individual teachers and the incentives principals face under Florida’s test-based accountability system.

The relative importance of intelligence, subject knowledge, and teaching skills in determining teacher productivity has important implications for recruiting and preparing future teachers. The apparent role of intelligence seems to suggest that policies designed to reduce entry barriers and encourage the “brightest” into the teaching profession could boost student achievement. However, this assumption is tempered by the fact that subject matter knowledge and teaching skills seem to matter as well. Sheer intelligence may not be enough; “good” teachers likely need to have adequate training in subject matter content and essential teaching techniques.

Our analysis of the predictive power of principal ratings and past value added also informs the current policy debate over the use of test scores and subjective evaluations to assess current teachers. Principal evaluations could be an important component of retention and tenure decisions if they either measure the same factors more precisely than do past value-added measures or if they encompass a broader range of factors that are important determinants of teacher productivity. We find some support for both these possibilities. When value-added measures are constructed from multiple years of test score data, past value added does a much better job at predicting future teacher performance than do principal evaluations. If one uses only a single year of information to estimate teacher value added, principal evaluations outperform past value added in predicting future teacher productivity. When a precise estimate of past value added is constructed from multiple years of data, principal ratings still add information that significantly improves the ability to predict future teacher performance in reading, but not in math. Many teachers, though, are relatively new to the job, and for these teachers precise estimation will always be a challenge. In addition, current merit pay plans for teachers commonly use only one year of data even when more years are available. The use of the principal evaluation might be particularly useful in these cases.

While this analysis is informative about the various ways to assess teachers, it is important to be cautious in drawing conclusions from these results for educational policies. For example, the fact that principals' assessments are positively related to value added and are sometimes better predictors of future value added than other indicators does not necessarily mean that evaluating teachers based on principals' assessments would be a wise policy. The assessments that principals offered in our study involved no financial or employment implications for teachers, and it is likely that the principals' stated judgments would differ in a high-stakes context. Also, even if principals were to give the same assessments in high-stakes settings, doing so could influence the relationship between principals and teachers in unproductive ways. Nevertheless, the fact that principal evaluations are better predictors of a teacher's contribution to student achievement than are traditional teacher credentials does not lend much support to current policies that reward teachers based on experience and formal education. The subjective principal ratings and objective value-added measures considered here are therefore worth considering as alternatives to the present system of teacher compensation.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. “Teachers and Student Achievement in the Chicago Public High Schools.” *Journal of Labor Economics* 25(1): 95–135.

Armor, David, Patricia Conry-Oseguera, Millicent Cox, Nicelma King, Lorraine McDonnell, Anthony Pascal, Edward Pauly, and Gail Zellman. 1976. *Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools*. Report #R-2007-LAUSD. Santa Monica, Calif.: RAND Corporation.

Bommer, William H., Jonathan L. Johnson, Gregory A. Rich, Philip M. Podsakoff, and Scott B. MacKenzie. 1995. “On the Interchangeability of Objective and Subjective Measures of Employee Performance: A Meta-analysis.” *Personnel Psychology* 48(3): 587–605.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. “The Economics and Psychology of Personality Traits.” *Journal of Human Resources* 43(4): 972–1059.

Borghans, Lex, Bas ter Weel, and Bruce A. Weinberg. 2008. “Interpersonal Styles and Labor Market Outcomes.” *Journal of Human Resources* 43(4): 815–58.

Boyd, Donald J., Pamela L. Grossman, Hamilton Lankford, Susanna Loeb, and Jim H. Wyckoff. 2008. “Overview of Measuring Effect Sizes: The Effect of Measurement Error.” Policy Brief 2. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education, Urban Institute.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. “Teacher-Student Matching and the Assessment of Teacher Effectiveness.” *Journal of Human Resources* 41:778–820.

———. 2007a. “How and Why Do Teacher Credentials Matter for Student Achievement?” Working Paper 2. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education, Urban Institute.

———. 2007b. “Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects.” Working Paper 11. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education, Urban Institute.

Cunha, Flavio, James Heckman, Lance Lochner, and Dimitry Masterov. 2006. “Interpreting the Evidence on Life Cycle Skill Formation.” In *Handbook of the Economics of Education*, ed. Eric A. Hanushek and Frank Welch, 697–812. Amsterdam: North-Holland.

Gallagher, H. Alix. 2004. “Vaughan Elementary’s Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement.” *Peabody Journal of Education* 79(4): 79–107.

Goldhaber, Dan. 2007. "Everyone's Doing It, but What Does Teacher Testing Tell Us about Teacher Effectiveness?" Working Paper 9. Washington, D.C.: National Center for the Analysis of Longitudinal Data in Education Research, Urban Institute.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." Discussion Paper 2006-01. Washington, D.C.: Brookings Institution.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24(3): 1141–77.

———. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19(2): 141–64.

Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin. 2005. "The Market for Teacher Quality." Working Paper 11154. Cambridge, Mass.: National Bureau of Economic Research.

Harcourt Assessment. 2002. "SAT-10 to SAT-9 Scaled Score to Scaled Score Conversion Tables." Unpublished computer tabulations. San Antonio, TX: Harcourt Assessment.

Harris, Douglas N., Carolyn D. Herrington, and Amy Albee. 2007. "The Future of Vouchers: Lessons from the Adoption, Design, and Court Challenges of Florida's Three Voucher Programs." *Educational Policy* 21(1): 215–44.

Harris, Douglas N., Stacey Rutledge, William Ingle, and Cynthia Thompson. Forthcoming. "Mix and Match: What Principals Really Look for When Hiring Teachers." *Education Finance and Policy*.

Harris, Douglas N., and Tim R. Sass. 2006. "Value-Added Models and the Measurement of Teacher Quality." Unpublished. Tallahassee, Fla.: Florida State University.

———. 2008. "Teacher Training, Teacher Quality and Student Achievement." Working Paper 3. Washington, DC: Center for the Analysis of Longitudinal Data in Education Research, Urban Institute.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24(3): 411–82.

Heneman, Robert L. 1986. "The Relationship between Supervisory Ratings and Results-Oriented Measures Performance: A Meta-analysis." *Personnel Psychology* 39:811–26.

Jacob, Brian A., and Lars Lefgren. 2005. "Principals as Agents: Subjective Performance Measurement in Education." Working Paper 11463. Cambridge, Mass.: National Bureau of Economic Research.

———. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics* 26(1): 101–36.

Jepsen, Christopher. 2005. “Teacher Characteristics and Student Achievement: Evidence from Teacher Surveys.” *Journal of Urban Economics* 57(2): 302–19.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2006. “What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City.” Working Paper 12155. Cambridge, Mass.: National Bureau of Economic Research.

Kimball, Steven M., Brad White, Anthony T. Milanowski, and Geoffrey Borman. 2004. “Examining the Relationship between Teacher Evaluation and Student Assessment Results in Washoe County.” *Peabody Journal of Education* 79(4): 54–78.

Koedel, Cory. 2009. “An Empirical Analysis of Teacher Spillover Effects in Secondary School.” Working Paper. Columbia, Mo.: University of Missouri-Columbia.

Koedel, Cory, and Julian Betts. 2009. “Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique.” Working Paper. Columbia, Mo.: University of Missouri-Columbia.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. Forthcoming. “The Intertemporal Variability of Teacher Effect Estimates.” *Education Finance and Policy*.

Medley, Donald M., and Homer Coker. 1987. “The Accuracy of Principals’ Judgments of Teacher Performance.” *Journal of Educational Research* 80(4): 242–47.

Mihaly, Kata, Daniel F. McCaffrey, J. R. Lockwood, and Tim R. Sass. 2009. “Centering and Reference Groups for Estimates of Fixed Effects: Modifications to Felsdvreg.” Unpublished manuscript.

Milanowski, Anthony T. 2004. “The Relationship between Teacher Performance Evaluation Scores and Student Assessment: Evidence from Cincinnati.” *Peabody Journal of Education* 79(4): 33–53.

Morris, Carl N. 1983. “Practical Empirical Bayes Inference: Theory and Applications.” *Journal of the American Statistical Association* 78(381): 47–55.

Murnane, Richard J. 1975. *The Impact of School Resources on the Learning of Inner City Children*. Cambridge, Mass.: Ballinger Publishing Company.

Podgursky, Michael J., and Matthew G. Springer. 2007. “Teacher Performance Pay: A Review.” *Journal of Policy Analysis and Management* 26(4): 909–49.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, Schools and Academic Achievement.” *Econometrica* 73(2): 417–58.

Rockoff, Jonah E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” *American Economic Review* 94(2): 247–52.

Rothstein, Jesse. 2009. “Teacher Quality in Educational Production: Tracking, Decay and Student Achievement.” Unpublished manuscript.

Strizek, Gregory A., Jayme L. Pittsonberger, Kate E. Riordan, Deanna M. Lyter, and Greg F. Orlofsky. 2006. *Characteristics of Schools, Districts, Teachers, Principals, and School Libraries in the United States: 2003-04 Schools and Staffing Survey*. NCES 2006-313 Revised. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Todd, Petra E., and Kenneth I. Wolpin. 2003. “On the Specification and Estimation of the Production Function for Cognitive Achievement.” *Economic Journal* 113(485): F3–33.

Varma, Arup, and Linda K. Stroh. 2001. “The Impact of Same-Sex LMX Dyads on Performance Evaluations.” *Human Resource Management* 40(4): 309–20.

CALDER