NBER WORKING PAPER SERIES

PRINCIPALS AS AGENTS: SUBJECTIVE PERFORMANCE MEASUREMENT IN EDUCATION

Brian A. Jacob
Lars Lefgren

Working Paper 11463
http://www.nber.org/papers/w11463

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2005

We would like to thank Joseph Price and J.D. LaRock for their excellent research assistance. We thank David Autor, Joe Doyle, Sue Dynarski, Amy Finkelstein, Chris Hansen, Robin Jacob, Jens Ludwig, Frank McIntyre, Jonah Rockoff, Doug Staiger, Thomas Dee and seminar participants at UC Berkeley, Northwestern, BYU, Columbia, Harvard, MIT and the University of Virginia for helpful comments. All remaining errors are our own. Jacob can be contacted at: John F. Kennedy School of Government, Harvard University, 79 JFK Street, Cambridge, MA 02138; email: brian_jacob@harvard.edu. Lefgren can be contacted at: Department of Economics, Brigham Young University, 130 Faculty Office Building, Provo, UT 84602-2363; email: l-lefgren@byu.edu.

Principals as Agents: Subjective Performance Measurement in Education
Brian A. Jacob and Lars Lefgren
NBER Working Paper No. 11463
June 2005
JEL No. I20, I28, J30, J70

ABSTRACT

In this paper, we compare subjective principal assessments of teachers to the traditional determinants of teacher compensation – education and experience – and another potential compensation mechanism -- value-added measures of teacher effectiveness based on student achievement gains. We find that subjective principal assessments of teachers predict future student achievement significantly better than teacher experience, education or actual compensation, though not as well as value-added teacher quality measures. In particular, principals appear quite good at identifying those teachers who produce the largest and smallest standardized achievement gains in their schools, but have far less ability to distinguish between teachers in the middle of this distribution and systematically discriminate against male and untenured faculty. Moreover, we find that a principal's overall rating of a teacher is a substantially better predictor of future parent requests for that teacher than either the teacher's experience, education and current compensation or the teacher's value-added achievement measure. These findings not only inform education policy, but also shed light on subjective performance assessment more generally.

Brian A. Jacob
JFK School of Government
Harvard University
79 JFK Street
Cambridge, MA 02138
and NBER
brian_jacob@ksg.harvard.edu

Lars Lefgren
Department of Economics
Brigham Young University
130 Faculty Office Building
Provo, UT 84602-2363
lars_lefgren@byu.edu

I. Introduction

Elementary and secondary school teachers in the U.S. traditionally have been compensated on the basis of experience and education according to formal salary schedules outlined in collective bargaining agreements. In recent years, however, many states have adopted accountability systems that reward and/or sanction schools, and in some cases even teachers, on the basis of student achievement scores (or gains). Many proposed merit pay programs go even further, linking teacher salaries directly to student achievement. Yet, recent studies have documented a number of undesirable consequences associated with such high-stakes testing policies, including teaching to the test and cheating (Jacob and Levitt, 2003, Jacob, 2005). A third option is to compensate teachers on the basis of principal evaluations. Subjective performance assessments not only play a large role in promotion and compensation decisions in many other occupations, but are currently used in education for the evaluation of untenured teachers.[1]

In this paper, we compare subjective principal assessments to the traditional determinants of teacher compensation – education and experience – and to value-added measures of teacher effectiveness based on student achievement gains. To do so, we assemble a unique data set that combines student achievement data with information on teachers including demographics, parent requests and principal ratings.

Because the success of a principal-based assessment system depends largely on which teacher qualities a principal values and whether a principal is able and willing to identify the teachers that demonstrate these qualities, we first examine how principals judge a teacher's

[1] In the current system, principals generally evaluate teachers each year, although their ability to fire or demote tenured teachers is quite limited. Principals maintain considerable control over the hiring of new teachers and the promotion of untenured teachers. Moreover, they have substantial informal power over tenured teachers through the assignment of various responsibilities and positions within the school (e.g., an after-school tutoring position that entails additional compensation).

ability to improve student performance as measured by standardized test scores. We find that principals are quite good at identifying those teachers who produce the largest and smallest standardized achievement gains in their schools (i.e., the top and bottom 10-20 percent), but have far less ability to distinguish between teachers in the middle of this distribution (i.e., the middle 60-80 percent). This is not a result of a highly compressed distribution of teacher ability, the lumpiness of the principal ratings or the differential precision of value-added measures across the distribution. However, we also find that principals systematically discriminate against male and untenured faculty.

To provide a more direct comparison between the three different methods of determining compensation – education and experience, principal ratings and value-added – we next investigate how well each of these measures predict student achievement and parent satisfaction. We find that subjective principal assessments of teachers predict future student achievement significantly better than teacher experience, education or actual compensation (which captures the combination of experience and education). This is true regardless of whether one uses a principal's rating of a teacher's ability to increase student achievement or a principal's overall rating of a teacher. While value-added measures of teacher effectiveness generally do a better job at predicting future student achievement than principal ratings, the two measures do almost equally well in identifying the best and worst teachers. These results suggest that student achievement (as measured by standardized test scores) would likely improve under a principal-based assessment system relative to the current system of compensation. To the extent that the most important staffing decisions involve sanctioning incompetent teachers and/or rewarding the best teachers, a principal-based system may also produce achievement outcomes roughly comparable to a test-based accountability system.

It is more difficult to ascertain how other important student outcomes might be affected if principal assessments played a larger role in promotion and compensation decisions. Assuming that parent preferences are a good overall measure of a wide variety of important student outcomes, an analysis of parent requests for teachers may shed light on this issue. We find that a principal's overall rating of a teacher is a substantially better predictor of future parent requests for that teacher than *either* the teacher's experience, education and current compensation *or* the teacher's value-added achievement measure.[2] This suggests that increasing a principal's ability to sanction and reward teachers would likely improve educational outcomes valued by parents but not readily captured by standardized tests.

There are at least two important caveats to interpreting our results. First, our analysis takes place in a context where principals were not explicitly evaluated on the basis of their ability to identify effective teachers.[3] It is possible that moving to a system where principals had more authority and responsibility for monitoring teacher effectiveness would enhance principals' ability to identify various teacher characteristics. On the other hand, it is possible that principals would be less willing to honestly assess teachers under such a system, perhaps because of social or political pressures. Second, our analysis focuses on the *source* of the teacher assessment; we do not address the type of rewards or sanctions associated with teacher performance. This is clearly an important dimension of any performance management system, and one would not expect either a principal-based or a test-based assessment system to have a substantial impact on student outcomes unless it were accompanied by meaningful consequences.[4]

[2] For a more detailed examination of parent preferences, see Jacob and Lefgren (2005).
[3] There were, however, a number of informal incentives for principals. For example, the district monitored the standardized test achievement of all schools and parents are an active presence in many of the schools.
[4] For examples of studies that examine accountability programs within education, see Jacob 2005, Kane and Staiger 2002, Figlio and Lucas 2004, Figlio and Winicki, forthcoming.

Our results suggest that switching from the current compensation system to one based on principal evaluations may improve both student achievement and parent satisfaction. While the relative merits of principal- versus test-based (i.e., value-added) assessment is less clear, principal assessment would likely enhance parent satisfaction and do relatively well at identifying those teachers who are particularly good or bad at promoting student achievement. Moreover, in theory principal-based evaluation has the potential to mitigate some of the negative consequences of test-based accountability. If principals can observe inputs as well as outputs, they may be able to ensure that teachers increase student achievement through improvements in pedagogy, classroom management or curriculum. Principals can also evaluate teachers on the basis of a broader spectrum of educational outputs.

More generally, our findings provide support for policies that invest principals with more authority for school-wide staffing decisions such as charter schools. The inability of principals to distinguish between a broad middle-range of teacher quality, however, suggests caution in relying on principals for fine grained performance determinations as might be required under certain merit pay policies.

In addition, our findings inform the education production function literature. A number of studies have documented substantial variation among teachers in their ability to raise student achievement (Murnane 1975, Hanushek 1992, Hanushek and Rivkin, 2005, Aaronson et. al. 2004, Rockoff 2004). Yet decades of research has found little association between teacher characteristics such as certification and student outcomes (Hanushek 1986, 1997).[5] A common perception among educators is that high quality teaching is like obscenity – it cannot easily be

[5] The one exception is experience early in a teacher's career. Several recent studies provide credible evidence that teacher quality increases with experience, at least for the first few years of a teacher's career (Rockoff 2004, Hanushek et. al. 2005).

defined, but one "knows it when [one] see[s] it."[6] Unfortunately, to date, there has been little evidence to confirm or disprove this hypothesis. Our results suggest that good teaching is, at least to some extent, observable by those close to the education process even though it may not be easily captured in those variables commonly available to the econometrician.

The paper also speaks to the broader literature on subjective performance assessment. While such evaluations are central to promotion, retention and compensation decisions in most industries, they have received relatively little attention in the economics literature (Prendergast 1999). We demonstrate the importance of accounting for estimation error in measured productivity when examining the correlation between subjective and objective measures of employee effectiveness. We also show that the relationship between subjective evaluations and actual productivity can vary substantially across the productivity distribution. Finally, we shed light on the extent and nature of discrimination in subjective evaluations.

The remainder of the paper proceeds as follows. In Section II, we review the literature on objective and subjective performance evaluation. In Section III, we describe our data and in Section IV outline how we construct the different measures of teacher effectiveness. The main results are presented in Section V. We conclude in Section VI.

II. Prior Literature on Performance Measurement

There is a long history of studies in organizational management, psychology and personnel economics that seek to determine the extent to which subjective supervisor ratings match objective measures of employee performance. Overall, this research suggests that there is a relatively weak relationship between subjective ratings and objective performance and that

[6] In 1964, Justice Potter Stewart tried to explain "hard-core" pornography, or what is obscene, by saying, "I shall not today attempt further to define the kinds of material I understand to be embraced . . . [b]ut I know it when I see it." Jacobellis v. Ohio, 378 U.S. 184, 197 (1964).

supervisor ratings are influenced by a number of demographic and interpersonal factors. In a meta-analysis of 23 studies of workers in a variety of jobs, Heneman (1986) found that supervisor ratings and objective performance were correlated 0.27 after correcting for sampling error and attenuation bias. A subsequent meta-analysis by Bommer et al. (1995) that included a larger percentage of sales jobs found a corrected mean correlation of 0.39. Several similar studies in education find a somewhat smaller correlation between principal-based teacher evaluations (Medley and Coker 1987, Peterson 1987, 2000). For example, Murnane (1975) and Armor et al. (1976) both found that principal evaluations of teachers predicted student achievement, even after conditioning on prior student test scores and a host of other student and classroom level demographic controls.[7] The authors of these studies argue that these results indicate that principals *can* identify effective teachers.[8]

The other main finding from this literature is that supervisor evaluations are often influenced by a number of non-performance factors such as the age and gender of the supervisor and subordinate and the likeability of the subordinate. For example, many studies document that that supervisors rate employees that they like, or perceive as similar to themselves, more highly than others, conditional on objective performance measures (Alexander & Wilkins, 1982; Bolino & Turnley, 2003; Heneman, Greenberger & Anonyuo, 1989; Lefkowitz, 2000; Wayne & Ferris,

[7] While it is difficult to directly compare these results to the education studies, the magnitude of the relationship appears to be modest. Murnane (1975) found that for third grade math, an increase in the principal rating of roughly 1 standard deviation was associated with an increase of 1.3 standard scores (or 0.125 standard deviations). The magnitude of the reading effect was somewhat smaller. Armor et al. (1976) found that a one standard deviation increase in teacher effectiveness led to a 1-2 point raw score gain (although it is not possible to calculate the effect size given the available information in the study).

[8] The few studies that examine the correlation between principal evaluations and other measures of teacher performance, such as parent or student satisfaction, find similarly weak relationships (Peterson 1987, 2000).

1990; Varma & Stroh, 2001). Prendergast (1999) observes that such biases create an incentive for employee's to engage in inefficient "influence" activities.[9]

While these findings are suggestive, this literature has many limitations. Most of the studies involve extremely small samples, often non-representative, samples. More importantly, they do not account for selection bias in constructing the objective productivity measures and do not adequately account for measurement error, which will tend to attenuate the estimated correlations.[10]

III. Data

The data for this study come from a mid-size school district located in the western United States.[11] The student data includes all of the common demographic variables as well as standardized achievement scores, and allows us to track the students over time. The teacher data, which we can link to students, includes a variety of teacher characteristics that have been used in previous studies, such as age, experience, educational attainment, undergraduate and graduate institution attended, and license and certification information. With the permission of the district, we surveyed all elementary school principals in February 2003 and asked them to rate the teachers in their schools along a variety of different performance dimensions.

To provide some context for the analysis, Table 1 shows summary statistics from the district. While the students in the district are predominantly white (73 percent), there is a reasonable degree of heterogeneity in terms of ethnicity and socioeconomic status. Latino

[9] Wayne and Ferris (1990) provide some empirical support for this hypothesis, finding that certain types of "influence tactics" such as ingratiation of the supervisor had a salutary effect on performance.
[10] In a seminal study of vocational rehabilitation counselors, for example, the researchers use the number of applications for service completed and the number of cases closed by the counselor as measures of objective productivity (Alexander and Wilkins 1982). To the extent that there is any non-random sorting of clients across rehabilitation centers or counselors, these measures are likely biased.
[11] The district has requested to remain anonymous.

students comprise 21 percent of the elementary population and nearly half of all students in the district (48 percent) receive free or reduced price lunch. Achievement levels in the district are almost exactly at the average of the nation (49th percentile on the Stanford Achievement Test).

The primary unit of analysis in this study is the teacher. To ensure that we could link student achievement data to the appropriate teacher, we limit our sample to elementary teachers who were teaching a core subject during the 2002-03 academic year.[12] We exclude kindergarten and first grade teachers because achievement exams are not available for these students.[13]

Our sample consists of 202 teachers in grades two through six. Like the students, the teachers in our sample are fairly representative of elementary school teachers nationwide. Only 16 percent of teachers in our sample are men. The average teacher is 42 years old and has roughly 12 years of experience teaching. The vast majority of teachers attended the main local university, while 10 percent attended another instate college and six percent attended a school out of state. 17 percent of teachers have a MA degree or higher, and the vast majority of teachers are licensed in either early childhood education or elementary education. Finally, 8 percent of the teachers in our sample taught in a mixed-grade classroom in 2002-03 and 5 percent were in a "split" classroom with another teacher.

In this district, elementary students take a set of "Core" exams in reading and math in grades 1 to 8.[14] These multiple-choice, criterion-referenced exams cover topics that are closely linked to the district learning objectives. While student achievement results have not been directly linked to rewards or sanctions until recently, the results of the Core exams are distributed

[12] We exclude non-core teachers such as music teachers, gym teachers and librarians.
[13] Achievement exams are given to students in grades one to six. In order to create a value-added measure of teacher effectiveness, it is necessary to have prior achievement information for the student, which eliminates kindergarten and first grade students.
[14] Students in select grades have recently begun to take a science exam as well. The district also administered the Standford Achievement Test (a national, norm-referenced exam) to students in grades three, five and eight over this period.

to parents and published annually. Citing these factors, district officials suggest that teachers and principals have focused on this exam even before the recent passage of the federal accountability legislation *No Child Left Behind*.

IV. Measures of Teacher Quality

Principal Assessments of Teacher Effectiveness

To obtain subjective performance assessments, we administered a survey to all elementary school principals in February 2003 asking them to evaluate their teachers along a variety of dimensions (see Appendix A for a sample survey form).[15] Principals were asked to rate teachers on a scale from 1 (inadequate) to 10 (exceptional). Importantly, principals were asked to not only provide a rating of overall teacher effectiveness, but also to assess a number of specific teacher characteristics including dedication and work ethic, classroom management, parent satisfaction, positive relationship with administrators and ability to raise math and reading achievement. Principals were assured that their responses would be completely confidential and would not be revealed to the teachers or to any other school district employee.

Table II presents the summary statistics of each rating. While there was some heterogeneity across principals, the ratings are generally quite high with an average of 8.07 and a 10-90 range from 6 to 10. The average rating for the least generous principal was 6.7. At the same time, however, there appears to be considerable variation within school. Figure I shows histograms where each teacher's rating has been normalized by subtracting the median rating

[15] In this district, principals conduct formal evaluations annually for new teachers and every third year for tenured teachers. However, prior studies have found such formal evaluations suffer from considerable compression with nearly all teachers being rated very highly. These evaluations are also part of a teacher's personnel file and it was not possible to obtain access to these without permission of the teachers.

within the school for that same item. It appears that principal ratings within school are roughly normally distributed with five to six relevant categories.

Because principal ratings differ in terms of the degree of leniency and compression, we normalize the ratings by subtracting from each rating the principal-specific mean for that question and dividing by the school-specific standard deviation. Perhaps not surprisingly, the principal responses to many (though not all) of the individual survey items are highly correlated (e.g., the correlation between teacher organization and classroom management exceeds 0.7 while the correlation between role model and relationship with colleagues is less than 0.4). To reduce the dimensionality of the principal ratings, we performed an exploratory factor analysis which yielded three factors.[16] Table III shows the factor loadings for the factors. The first factor clearly measures student satisfaction, with high loadings on principal ratings of student satisfaction and teacher as role model. The second factor appears to capture what might be described as traditional teaching ability, with high loadings on classroom management, organization and ability to improve student test scores. The third factor captures a teacher's collegiality, with high loadings on the items that ask principals to assess the teacher's relationship with colleagues and administrators.

Value-Added Measures of Teacher Ability to Raise Standardized Achievement Scores

The primary challenge to estimating measures of teacher effectiveness using student achievement data involves the potential for non-random assignment of students to classes.

[16] Because the principal evaluation of parent satisfaction may be highly correlated with the parent request measure that is included in some models, we exclude this item in creating the principal factors. While this increases the significance of the parental request measure, it does not impact any of the other estimates in the model. As an additional check, we create a second set of principal measures that are purged of the parent satisfaction information by regressing the factors created above on the parental satisfaction item. We then use the residuals from these regressions as factors that are by construction orthogonal to the principal's view of parent satisfaction. Aside from increasing the significance of the parent request measures, the results from using these factors are comparable to the results based on the original factors.

Following the standard practice in this literature, we estimate value-added models that control for a wide variety of observable student and classroom characteristics including prior achievement measures (see, for example, Aaronson et al. 2004, Rockoff 2004 and Hanushek and Rivkin 2004). The richness of our data allows us to observe teachers over multiple years, and thus to distinguish permanent teacher quality from idiosyncratic class-year shocks and to estimate a teacher experience gradient utilizing variation within individual teachers. In this section, we outline our general strategy for creating value-added measures. Appendix D presents a more detailed description of our approach and Appendix E shows that our main results are robust to a variety of alternative specifications.

For our baseline specification, we estimate the following model:

$$(1) \qquad y_{ijkt} = C_{jt}\mathrm{B} + X_{it}\Gamma + \psi_t + \phi_k + \delta_j + \alpha_{jt} + \varepsilon_{ijkt}$$

where *i* indexes students, *j* indexes teachers, *k* indexes school, and *t* indexes year. The outcome measure, y , represents a student's score on a math or reading exam. The scores are reported as the percentage of items the student answered correctly, but we normalize achievement scores to be mean zero and standard deviation one within each year-grade.

The vector X consists of the following student characteristics: age, race, gender, free-lunch eligibility, special education placement, limited English proficiency status, prior math achievement, prior reading achievement, and grade fixed effects. C is a vector of classroom measures that include indicators for class size and average student characteristics. ψ_t and ϕ_k are a set of year and school fixed effects respectively. Teacher *j*'s contribution to value-added is captured by the δ_j's . α_{jt} is an error term that is common to all students in teacher *j*'s classroom in period *t* (e.g., adverse testing conditions faced by all students in a particular class such as a barking dog). ε_{ijkt} is an error term that takes into account the student's idiosyncratic error. In

order to account for the correlation of students within classrooms, we correct the standard errors using the using the method suggested by Moulton (1990).[17]

While there is no way (short of randomly assigning students and teachers to classrooms) to completely rule out the possibility of selection bias, several pieces of evidence suggest that such non-random sorting is unlikely to produce a substantial bias in our case. First, to account for unobservable, time-invariant student characteristics student (e.g., motivation or parental involvement), we estimate value-added models that include student fixed effects λ_i with either achievement levels or gains as the dependent variable: [18]

$$(2) \qquad y_{ijkt} = C_{jt}\mathrm{B} + X_{it}\Gamma + \psi_t + \phi_k + \delta_j + \alpha_{jt} + \lambda_i + \varepsilon_{ijkt}$$

$$(3) \qquad y_{ijkt} - y_{ijkt-1} = C_{jt}\mathrm{B} + X_{it}\Gamma + \psi_t + \phi_k + \delta_j + \alpha_{jt} + \lambda_i + \varepsilon_{ijkt}$$

As we show in Appendix E, our results are robust to using value-added measures from specifications (2) and (3). The reason that we do not use value-added measures that include student fixed effects as our baseline specification is that this would require that we drop teachers who began teaching in our last year of data 2002-03.[19]

Second, in order to determine exactly how students are assigned to classrooms and to explicitly examine how the assignment process may influence our estimates, we conducted detailed interviews with the school principals. With the exception of certain schools and grades where students are tracked by ability for math, a "homeroom" teacher provides instruction in all

[17] Another possibility would be to use cluster-corrected standard errors. However, such standard errors cannot be computed for teachers that appear in the sample for a single year. Additionally, the estimated standard errors can behave very poorly for teachers that are in the sample for a small number of years. It is also possible to estimate a model that includes a random teacher-year effect, which should theoretically provide more efficient estimates. In practice, however, the random effect estimates are comparable to those we present in terms of efficiency and are considerably more difficult to estimate from a computational perspective. The intra-class correlation coefficients calculated as part of the Moulton procedure are roughly .06 in reading and .09 in mathematics.

[18] Note that in specification (2) the covariates include lagged achievement measures. In specification (3) lagged achievement measures are not included.

[19] If we include student fixed effects in a model that uses gains as the dependent variable, we cannot obtain estimates for sixth grade teachers who began teaching in 2002-03.

core subjects.[20] When assigning students to homeroom teachers, principals attempt to create classrooms that are balanced along a variety of dimensions, most notably race, gender and prior academic achievement. In all schools, however, parents are able to request a particular teacher and most principals indicate that they attempt to honor such requests when this does not interfere with the goal of creating balanced classrooms. Value-added models that include student fixed effects will control for any permanent, unobserved quality that may be correlated with parent requests and achievement. But any unobserved, time-varying factors correlated with requests and achievement may still generate a bias in our estimates. In order to investigate this possibility, we compared the achievement trends of children whose parents submitted requests with their peers. Conditional on initial achievement and basic demographics, we find that the students whose parents submit requests do not perform significantly better or worse than non-requesting students. This suggests that teacher assignment on the basis of parent requests is unlikely to be highly correlated with *unobserved* student ability.[21]

The second major concern in estimating value-added measures of teacher quality involves estimation error, which arises not only from sampling variation, but also from idiosyncratic factors that operate at the classroom level in a particular year (e.g., a dog barking in the playground, a flu epidemic during testing week, or something about the dynamics of a particular group of children).[22] Measurement error will lead us to understate the correlation between the principal ratings and the value-added indicators. Note that the observed value-added can be written as $\hat{\delta}^{OLS} = \delta + e$ where δ is the true fixed effect and e represents estimation error.

[20] For more detail about math tracking, and how this is handled in the value-added models, see Appendix B.
[21] This analysis is based on a single school where we have individual level parent request data. There is no reason, however, to believe that parents making requests in this school differ in unobserved ways than parents making a request in some other school.
[22] We will use the terms estimation error and measurement error interchangeably, although in the testing context measurement error often refers to the test-retest reliability of an exam whereas the error stemming from sampling variability is described as estimation error.

If we denote the principal rating as $\hat{\delta}^P$, then it is simple to show that the correlation between principal rating and *observed* value-added is biased downward relative to the correlation between principal rating and *true* value-added:

$$(4) \qquad Corr\left(\hat{\delta}^p, \hat{\delta}^{OLS}\right) = \frac{Cov\left(\hat{\delta}^p, \hat{\delta}^{OLS}\right)}{\sqrt{Var\left(\hat{\delta}^P\right)Var\left(\hat{\delta}^{OLS}\right)}} = \frac{Cov\left(\hat{\delta}^p, \delta\right)}{\sqrt{Var\left(\hat{\delta}^P\right)\left[Var(\delta)+Var(e)\right]}} < Corr\left(\hat{\delta}^p, \delta\right)$$

Fortunately, it is relatively simple to correct for this using the observed estimation error from the value-added model. We obtain a measure of the true variance by subtracting the mean error variance (the average of the squared standard errors on the estimated teacher fixed effects) from the variance of the observed valued-added measures: $Var(\delta) = Var\left(\hat{\delta}^{OLS}\right) - Var(e)$.[23]

Then we can then simply multiply the observed correlation, $Corr\left(\hat{\delta}^p, \hat{\delta}^{OLS}\right)$, by $\frac{\sqrt{Var\left(\hat{\delta}^{OLS}\right)}}{\sqrt{Var(\delta)}}$ to obtain the adjusted correlation.[24] We obtain the standard errors using a bootstrap.[25]

In addition to biasing our correlations, estimation error will lead to attenuation bias if we use the teacher value-added measures as an explanatory variable in a regression context.[26] To

[23] This assumes that the OLS estimates of the teacher fixed effects are not correlated with each other. This would be true if the value-added estimates were calculated with no covariates. Measurement error in the coefficients of the covariates generates a non-zero covariance between teacher fixed effects, though in practice the covariates are estimated with sufficient precision that this is not a problem.

[24] This adjustment assumes that principal's rating is unrelated to the error of our OLS estimate of teacher effectiveness. Specifically, we assume that the numerator in equation (4) can be rewritten as follows: $Cov\left(\hat{\delta}^p, \hat{\delta}^{OLS}\right) = Cov\left(\hat{\delta}^p, \delta\right) + Cov\left(\hat{\delta}^p, e\right)$. This would not be true if the principals were doing the same type of statistical analysis as we are to determine teacher effectiveness. However, to the extent that principals base their ratings largely on classroom observations, discussions with students and parents and other factors unobservable to the econometrician, this assumption will hold. To the extent that this is not true and principals do base their ratings solely on the observed test scores (in the same manner as the value-added model does – that is, conditioning on a variety of covariates), the correlation we calculate will be biased upwards.

[25] In the bootstrap, we stratify by school, consistent with the thought experiment of randomly drawing new teachers for each principal.

[26] If the value-added measure is used as a dependent variable, it will lead to less precisely estimated estimates relative to using a measure of true teacher ability. Measurement error will also lead us to overstate the variance of teacher effects, although this is a less central concern for the analysis presented here.

account for attenuation bias when we use the teacher value-added in a regression context, we construct empirical Bayes (EB) estimates of teacher quality. This approach was suggested by Kane and Staiger (2002) for producing efficient estimates of school quality, but has a long history in the statistics literature (see, for example, Morris, 1983).[27] Appendix C describes this procedure in detail and formally illustrates that it eliminates attenuation bias.

Before we turn to our primary objective, it is useful to consider the teacher value-added measures that we estimate. After adjusting for estimation error, we find that the standard deviation of teacher quality is 0.19 in reading and 0.32 in math, which appears roughly consistent with recent literature on teacher effects.[28] Because the dependent variable is a state-specific, criterion referenced test that we have normalized within grade-year for the district, in order to provide a better sense of the magnitude of these effects, we take advantage of the fact that in recent years third and fifth graders in the district have also taken the nationally normed Stanford Achievement Test (SAT9) in reading and math so that one can determine how a one standard deviation unit change on the Core exam translates into national percentile points. This comparison suggests that moving a student from an average teacher to a teacher one standard deviation above the mean would result in roughly a 4-5 percentile point increase in test scores. Because of the non-linearity of the scales, a move from the average teacher to a teacher two standard deviations above the mean in terms of value-added would result in an increase of over

[27] In fact, the EB approach described here is very closely related to the errors-in-variables approach that allows for heteroskedastic measurement error outlined by Sullivan (2001).

[28] Hanushek et al. (2005), for example, find that one standard deviation in the teacher quality distribution is associated with a 0.22 standard deviation increase in math on the Texas state assessment. Rockoff (2004) finds considerably smaller effects – namely that a one standard deviation increase in the teacher fixed effect distribution raises student math and reading achievement by roughly 0.10 standard deviations on a nationally standardized scale. This is likely for two reasons. First, 0.1 standard deviations on a national scale (which Rockoff uses) may be much more than 0.1 standard deviations within the distribution of achievement within Rockoff's districts. This would be true if the students in his district were substantially more homogenous than is true nationally. Since our study uses the within-district distribution as opposed to the national distribution of student achievement, our results are not directly comparable to Rockoff's. In addition, teachers in Rockoff's district might be more homogenous than is true in our sample. To the extent that Rockoff's districts serve more affluent towns with fewer poor neighborhoods, this explanation might be plausible.

12 percentile points. The range of the 95 percent confidence interval around the mean teacher quality in the district is roughly 22 percentile points. Given that the average student in the district scores at the 49th percentile, this suggests that there is quite considerable variation in teacher quality in the district.

V. Can Principals Identify Effective Teachers?

The success of a principal-based assessment system depends largely on which teacher qualities a principal values and whether a principal is able and willing to identify the teachers that demonstrate these qualities. This section presents results on three key empirical questions: (1) How well can principals identify teacher effectiveness at raising student achievement? (2) What teacher characteristics do principals value? (3) Do principals discriminate against certain teachers? In the following section, we directly compare principal assessments to the traditional determinants of teacher compensation (i.e., education and experience) and to value-added measures of teacher quality that are based on student achievement gains.

Can Principals Identify a Teacher's Ability to Raise Standardized Test Scores?

To begin answering our first question, Table IV shows the correlation between a principal's subjective evaluation of how effective a teacher is at raising student reading (math) achievement and that teacher's actual ability to do so as measured by the value-added measures described in the previous section. Columns 1 and 3 (of row 1) show unadjusted correlations of 0.20 and 0.28 for reading and math respectively. As discussed earlier, however, these correlations will be biased toward zero because of the estimation error in the value-added measures.

Once we adjust for estimation error, the correlations for reading and math increase to 0.32 and 0.36 respectively. It is important to emphasize that these correlations are based not on a principal's overall rating of the teacher, but rather on the principal's assessment of how effective the teacher is at "raising student math (reading) achievement." Because the subjective and objective measure are identifying the same underlying construct, they should not be biased downward as in the case with many prior studies of subjective performance evaluation.[29] The positive and significant correlations indicate that principals do have some ability to identify this dimension of teacher effectiveness. These basic results are robust to a wide variety of alternative specifications and sensitivity analyses described in detail in Appendix E.

However, one might ask why these correlations are not even higher. One possibility is that principals focus on the average test scores in a teacher's classroom rather than student *improvement* relative to students in other classrooms. The correlations between principal ratings and average student achievement scores by teacher, shown in row 2, provide some support for this hypothesis. The correlation between principal ratings and average test scores in reading is significantly higher (0.56 versus 0.32) than the correlation with between principal ratings and teacher value-added. This suggests that principals may base their ratings at least partially on a naïve recollection of student performance in teacher's class. Another reason may be that principals focus on their most recent observations of teachers. In results not shown here, we find

[29] Bommer, Johnson, Rich, Podsakoff, et al. (1995) emphasize the potential importance of this issue, noting that in the three studies they found where objective and subjective measures tapped precisely the same performance dimension, the mean corrected correlation was 0.71 as compared with correlations of roughly 0.30 in other studies. Medley and Coker (1987) are unique in specifically asking principals to evaluate a teacher's ability to improve student achievement. They find that the correlation with these subjective evaluations are no higher than with an overall principal rating.

that the average achievement score (or gains) in a teacher's classroom in 2002 is a significantly stronger predictor of the principal's rating than the scores (or gains) in any prior year.[30]

While the results provide evidence that principals have some ability to evaluate teacher ability, it is difficult to know whether these correlations are in fact large or small.[31] Correlations are not only quite sensitive to outliers, but it is also not clear what scale the principals are using to assess teachers. Finally, a simple correlation does not tell us whether principals are more effective at identifying teachers at certain points on the ability distribution. For these reasons, we turn to a non-parametric measure of the association between ratings and productivity.[32]

Table V shows the estimates of the percent of teachers that a principal can correctly identify in the top (bottom) group within his or her school. If we knew the true ability of each teacher, this exercise would be trivial. In order to address this question using our noisy measure of teacher effectiveness, we rely on a Monte Carlo simulation in which we assume that a teacher's true value-added is distributed normally with a mean equal to the point estimate of the teacher fixed effect and a standard deviation equal to the standard error on the teacher's estimate.

[30] Of course, it is possible that principals may be correct in assuming that teacher effectiveness changes over time so that the most recent experience of a teacher may be the best predictor of actual effectiveness. To examine this possibility, we create value-added measures that incorporate a time-varying teacher experience measure. As shown in Table E1, we obtain comparable results when we use this measure.

[31] One possible caveat throughout the analysis is that the lumpiness of the principal ratings reduces the observed correlation between principal ratings and actual value-added. To determine the possible extent of this problem, we performed a simulation in which we assumed principals perfectly observed a normally distributed teacher quality measure. Then the principals assigned teachers in order to the actual principal reading rankings. For example, a principal who assigned 2 6's, 3 7's, 6 8's, 3 9's, and 1 10, would assign the two teachers with the lowest generated value-added measures a 6. She would assign the next three teachers 7's and so on. The correlation between the lumpy principal rankings and the generated teacher quality measure is about 0.9, suggesting that at most the correlation is downward biased by about 0.1 due to the lumpiness. When we assume that the latent correlation between the principal's continuous measure of teacher quality and true effectiveness is 0.5, the correlation between the lumpy ratings and the truth is biased downwards by about 0.06, far less than would be required to fully explain the relatively low correlation between the principal ratings and the true teacher effectiveness. In practice, the bias from lumpiness is likely to be even lower. This is because teachers with dissimilar quality signals are unlikely to be placed in the same category—even if no other teacher is between them. In other words, the size and number of categories is likely to reflect the actual distribution of teacher quality, at least in the principal's own mind.

[32] The correlations (and associated non-parametric statistics) may understate the relation between objective and subjective measures if principals have been able to remove or counsel out the teachers that they view as the lowest quality. However, our discussions with principals and district officials suggest that this occurs rarely and is thus unlikely to introduce a substantial bias in our analysis.

The basic intuition is that by taking repeated draws from the value-added distribution of each teacher in a school, we can determine the probability that any particular teacher will fall in the top or bottom group within his or her school, which we can then use to create the conditional probabilities shown in Table V. (Appendix D provides a more detailed discussion of these calculations.)

Examining the results in the top panel, we see that the teachers identified by principals as being in the top category were, in fact, in the top category according to the value-added measures about 52 percent of the time in reading and 69 percent of the time in mathematics. If principals randomly assigned ratings to teachers, we would expect the corresponding probabilities to be 14 and 26 percent respectively. This suggests that principals have considerable ability to identify teachers in the top of the distribution. The results are similar if one examines teachers in the bottom of the ability distribution (bottom panel).

The second and third panels in Table V suggest that principals are significantly *less* successful at distinguishing between teachers in the middle of the ability distribution. For example, in the second panel we see that principals correctly identify only 49 percent of teachers as being better than the median, relative to the null hypothesis of 33 percent – i.e., the percent that one would expect if principal ratings were randomly assigned. The difference of 16 percentage points is considerably smaller than the difference of 38 percentage points we find in the top panel. There is a similar picture at the bottom of the distribution. Principals appear somewhat better at distinguishing between teachers in the middle of the math distribution compared with reading, but they again appear to be better at identifying the best and worst teachers.

One reason that principals might have difficulty distinguishing between teachers in the middle is that the distribution of teacher value-added is highly compressed. However, the box plots shown in Figure II suggest that this is not the case. Teachers who receive principal ratings at or close to the median in the school have estimated value-added measures that are quite widely dispersed. In contrast, nearly all of the teachers in the top (bottom) principal categories have estimated value-added measures (including measurement error) that place them above (below) their school average. Figure III shows scatterplots and lowess lines of the relationship between principal ratings and estimated teacher value-added measures for the entire sample and the sample of teachers who did not receive the top or bottom principal rating (in the bottom panel). If we exclude those teachers who received the best and worst ratings, the adjusted correlation between principal rating and teacher value-added is 0.06 and -0.02 in reading and math respectively (neither of which is significantly different than zero). These results suggest that principals may not be sufficiently familiar with the educational production function to identify the subtle differences in instructional style that lead to marginally different student outcomes.

What Teacher Characteristics do Principals Value?

The implications for moving to a system of compensation based on principal assessment depends not only on how well principals can identify effective teachers, but also on the relative importance they place on a teacher's ability to raise standardized test scores. While such preferences could theoretically be set by district administrators or other policymakers, it is likely that principals would retain some autonomy over personnel decisions so their preferences are important to investigate. Because principals were asked to provide an overall rating of each

teacher as well as an assessment of a number of specific teacher attributes, it is possible to examine how principals value each of these dimensions of teacher quality.

Table VI shows the results from a number of different OLS regressions in which the dependent variable is always the principal's overall rating of the teacher. As noted earlier, because many of the individual survey items are highly correlated with each other, we focus on the composite factors described in the previous section. Column 1 shows how much each of the three factors – achievement, collegiality and student satisfaction – contributes to a principal's overall evaluation of a teacher. While all three factors are positively related to the overall rating, the achievement factor is the most important predictor. A one standard deviation increase in the a principal's evaluation of a teacher's management and teaching ability, for example, is associated with a 0.56 standard deviation increase in the principal's overall rating.[33]

Columns 2 and 3 show that parental requests are significantly related to the overall principal evaluation, both independently and conditional on the three factors. Columns 4 and 5 show the value-added measures are correlated with the overall principal rating. While several observable teacher characteristics are significantly related to the principal's overall rating (column 6), once we control for the principal factors these relationship disappear.

Do Principals Discriminate?

Prior literature suggests that subjective performance evaluations may be biased. To the extent that this is true in the context of principals and teachers, a move toward principal assessment would have important efficiency as well as equity implications since teachers may

[33] In results not shown here, we find that the preference for student achievement indicated by the large positive coefficient on the achievement factor stems primarily from the student achievement and classroom organization management items and less from the work ethic item. Interestingly, principals seem to place equal weight on math and reading achievement. They also place similar weight on organization and management items as they do on the achievement items.

have an incentive to engage in unproductive activities in order to improve their subjective evaluation. In most empirical studies of discrimination, one of the greatest challenges is to accurately measure an employee's true productivity. If this measure is omitted from the analysis of wage differentials, for example, it is difficult to interpret the residual in a wage regression as stemming from discrimination as opposed to unobserved (to the researcher) productivity. The availability of objective as well as subjective performance measures allows us to overcome this hurdle.

We will thus define discrimination as the practice whereby principals give systematically lower ratings to a specific group of teachers holding constant actual productivity. Specifically, we estimate the following OLS regression

$$(5) \qquad \hat{\delta}_j^P = \alpha_0 + \alpha_1 \delta_j + \mathrm{A} X_j + e_j$$

where X is a vector of teacher characteristics and the other measures are defined as before. Importantly, note that the outcome measure is the principal's assessment of a teacher's ability to raise reading (math) scores rather than the principal's overall assessment of the teacher. While a teacher's true ability to raise student achievement, δ_j, is not observed, using the EB estimate of teacher value-added eliminates attenuation bias and recovers consistent estimates of all parameters.[34]

Table VII presents the results from estimating equation (5). Columns 1 and 4 present baseline estimates that do not control for teacher value-added. Here we see that male and untenured teachers receive significantly lower ratings than their female and tenured counterparts. In columns 2 and 5, we control for EB estimates of the teacher value-added and the results remain largely unchanged. Specifically, principals rate both male and untenured teachers

[34] Mathematically, an error-in-variables regression employs a procedure that is virtually identical to the shrinkage used to construct our EB measure of teacher effectiveness.

roughly 0.3 to .5 standard deviations lower than their female and tenured colleagues with the same actual proficiency.[35]

These results imply that principals discriminate against male and untenured teachers in the assignment of the achievement ratings. But what might explain this behavior? The economics literature on racial and gender discrimination in the workplace often distinguishes between preference-based and statistical discrimination. In the first case, employers discriminate against a particular group because of a dislike for that group. In the second case, the employer uses information on the distribution of productivity across groups as a proxy for individual productivity in the absence of perfect information on the individual. A third potential motivation for discrimination might be described as incorrect statistical discrimination. In this case, employers mistakenly believe that certain groups are less productive than others. While we cannot definitively ascertain the motivation behind the differential ratings we observe, a closer examination of principal behavior sheds some light on the nature of the discrimination.

Several pieces of evidence suggest that the observed discrimination is not preference-based. First, male and female principals both rate male teachers lower than female teachers, suggesting that a gender-specific bias is not likely to be driving the results. Second, controlling for the principal's self-reported relationship with the teacher (in columns 3 and 6 of Table VII) does not change the results, as one would expect if the negative coefficients on male and untenured were driven by the fact that principals simply dislike teachers who are male or untenured (or teachers with characteristics that are correlated with being male or untenured).[36] Interestingly, however, principals rate "favored" teachers more highly than others – e.g., teachers

[35] The inclusion of the other principal measures and parent request measure shown in Table VII does not affect the coefficients on male and untenured either.

[36] Naturally, this presupposes that a principal's relationship with a specific teacher is a good proxy for the degree of prejudice felt toward a specific individual.

that are one standard deviation higher on the principal relationship scale score roughly one-third of a standard deviation higher on principal rating. Again, to the extent that this bias provides an incentive for non-productive influence activity on the part of teachers, it may reduce the performance of the school overall.[37] Finally, the fact that we do not see similar negative effects for males and untenured teachers when we examine the principal's overall rating in Table VI (particularly once we condition on the component principal measures) suggests that the gender and experience discrimination manifests itself only on principal ratings regarding a teacher's ability to increase student test performance. If animus were the underlying source of discrimination, we might expect males and untenured teachers to perform worse on the overall rating as well.

Next consider the possibility of statistical discrimination. At first glance, this may seem to be a plausible explanation. There is a widespread belief among educators that new teachers perform worse than experienced teachers, and recent evidence supports this perception (Rockoff 2004 and Hanushek et al. 2005). While there is not similar evidence regarding gender effects, there is a perception among some that women are more effective teachers of young children than men because of certain personality traits. And, in fact, we find some evidence for both effects in our data. Table VIII presents OLS estimates of the relationship between teacher characteristics and estimated value-added. Although the statistical power is quite low, the point estimates suggest that male teachers are roughly 0.4 standard deviations (on the teacher quality distribution) worse than female teachers in reading and that untenured teachers are about 0.33 standard deviations worse than tenured teachers in math.

[37] It is worth noting that student satisfaction (along with male, untenured and principal relationship with the teacher) are all significantly related to principal ratings in bivariate regressions.

However, the fact that the inclusion of value-added does not affect the estimated impact of the male and untenured indicator variables on principal ratings suggests that principals may simply have a misguided view of the education production function and/or the characteristics associated with gender and tenure. For example, principals may believe (correctly or incorrectly) that the teacher's ability to raise student test scores is a function of empathy and that the female teachers in his or her school have more empathy than the male teachers.[38] Regardless of the cause, however, this discrimination may place male and untenured teachers at a disadvantage in a system that relies more heavily on principal assessment.

VI. A Comparison of Alternative Teacher Assessment Measures

In this section, we directly compare principal assessments to the traditional determinants of teacher compensation – education and experience – as well as value-added measures of teacher quality that are based on student achievement gains. Specifically, we examine how well each of the three proxies for teacher quality – compensation, principal assessment and estimated value-added – predict student achievement and parent requests.[39]

Student Achievement

In order to examine how well each of the teacher quality measures predict student achievement, we regress 2003 math and reading scores on prior student achievement, student

[38] Alternatively, these results may still be based on rational statistical discrimination if principals are largely unaware of a teacher's actual effectiveness and thus base their ratings almost exclusively on average productivity differences between groups. This explanation is at odds with the principals' demonstrated ability to identify effectiveness at the extremes of the distribution. It is a plausible source of discrimination, however, toward the majority of teachers in the middle of the performance distribution.

[39] Note that when comparing the predictive power of the various measures, we are essentially comparing the principal and compensation measures against *feasible* value-added measures. Using unobserved actual value-added could increase the predictive power (as measured by the r-squared), but this is not a policy relevant measure of teacher quality. Of course, the nature of the EB measures is such that coefficient estimates are consistent measures of impact of actual teacher value-added.

demographics, and a set of classroom covariates including average classroom demographics and prior achievement and class size. We then include different measures of teacher quality. The standard errors shown account for the correlation of errors within classroom. Importantly, the value-added measures are calculated using the specification described in equation (1) but only include student achievement data from 1998 to 2002. In order to account for attenuation bias in the regressions, we use Empirical Bayes estimates of the value-added (see Section IV and Appendix C). It is important to note that the use of value-added measures based on 1998-2002 means that we cannot include any teachers with only 2003 student achievement data, which means that first-year teachers will be excluded in the subsequent analysis, which limits our sample to 160 teacher observations for reading and 116 teacher observations for math.[40] To make the coefficients comparable between the principal ratings and value-added, we divide each EB measure by the standard deviation of the EB measure itself. Thus the coefficient can be interpreted as the effect of moving one standard deviation in the *empirical* distribution of teacher quality.[41]

Table IX presents the results. Column 1 shows the effect of teacher experience and education on reading achievement. While there does not appear to be any significant relationship between teacher experience and student achievement, results not presented here indicate that this is due to the necessary omission of first-year teachers, who perform worse on average than experienced teachers.[42] In contrast, teachers with advanced degrees have students

[40] The sensitivity analysis in Table E1 indicates that excluding first-year teachers does not affect the estimated correlation between the value-added measures and principal ratings, which suggests this exclusion should not bias our results in this regard.

[41] Since we are comparing the relative value of using a test-based vs. principal-based measure of performance, the most relevant comparison is between a movement in the empirical (not actual) distribution of teacher effectiveness and the principal rating.

[42] The results in Table VIII suggest that untenured teachers have lower value-added measures than tenured teachers, and results from some of the alternative specifications discussed in Appendix E indicate that, conditional on teacher fixed effects, first-year teachers produce smaller achievement gains than more experienced teachers.

that score roughly 0.10 standard deviations higher than their counterparts (although this relationship should not be interpreted as causal since education levels may well be associated with other omitted teacher characteristics). In this district, however, compensation is a complicated, non-linear function of experience and education. Column 2 shows that actual compensation has no significant relationship to student achievement. In results not shown here, we find that including polynomials in compensation does not change this result. Columns 3 and 4 indicate that principal ratings – both overall ratings and ratings of a teacher's ability to raise achievement – are significantly associated with higher student achievement. Conditional on prior student achievement, demographics and classroom covariates, students whose teachers receive an overall rating one standard deviation above the mean are predicted to score roughly 0.06 standard deviations higher than students whose teacher received an average rating. Column 5 shows that a teacher's value-added is an even better predictor of future student achievement gains, with a coefficient almost twice as large as that on the principal ratings.[43] The r-square measures in the bottom row indicate that none of the measures explain a substantial portion of the variation across students, as one would expect given that much of the variation in nearly all student-level regressions occurs within classroom. Nonetheless, bootstrap tests indicate that the value-added measure explains significantly (at the 10 percent level in reading and the 5 percent level in math) more of the variation in student achievement than the principal ratings. As shown in columns 7-11, the results for math are comparable to the reading results.

To the extent that principal ratings are picking up a different dimension of quality than the test-based measures, one might expect that combining principal and value-added measures

[43] We examined the functional form of the relationship between both teacher quality measures and student achievement, but found that both were approximately linear. Also, when we do not normalize the EB measure by the standard deviation of teacher ability, the coefficient is insignificantly different from 1, which we would expect given that the EB is essentially the conditional expectation of teacher effectiveness.

would yield a better predictor of future achievement. Column 6 suggests that this might be the case. Conditional on teacher value-added, the principal's overall rating of a teacher is a significant (at the 10 percent level) predictor of student achievement. The results for math, shown in column 12, are even stronger.

Parent Satisfaction

It is more difficult to ascertain how other important student outcomes might be affected if principal assessments played a larger role in promotion and compensation decisions. To the extent that parent preferences are a good overall measure of a wide variety of important student outcomes, however, parent requests provide another way to compare the different assessment systems. Table X presents OLS estimates of the relationship between different measures of teacher quality and parent requests. The unit of observation is a teacher*year*grade combination and the sample includes elementary teachers who (a) have principal ratings and value-added measures and (b) taught grades 2-6 in our sample of schools in 2003-04 or 2004-05.[44] The final sample includes 213 observations, representing 144 different teachers.[45] The outcome measure is the number of parent requests that the teacher received in that year. To account for heteroskedasticity in parent requests due to school size and variation in school request policies, we normalize the number of requests by subtracting the average number of requests for a particular school*grade*year and dividing by the number of students in the cohort (i.e., the total number of parents who could have made a request). Our estimates will thus capture how particular teacher characteristics influence the percent of parents requesting a particular teacher. We include fixed effects for school*grade*year to ensure that our identification comes from

[44] For a more detailed description of the sample, see Jacob and Lefgren (2005).
[45] In the specifications in which we include math value-added measures, the sample falls to 156 teacher*year*grade combinations including 107 unique teachers

variation *within* the relevant choice set facing parents. Finally, we cluster-correct the standard errors to account for the fact that we observe certain teachers more than once.

The results indicate that a principal's overall rating of a teacher is a substantially better predictor of future parent requests for that teacher than *either* the teacher's experience, education and current compensation *or* the teacher's value-added achievement measure. The specification in column 1 includes indicators of teacher experience and education level, which together determine compensation. We see that the number of requests generally falls with experience, though our choice of sample does not allow us to observe requests for the newest teachers, for whom we have neither principal evaluations nor EB performance measures.[46] Teachers with an advanced degree do not receive more requests than colleagues who only have a BA degree. Column 2 shows that actual compensation (which is a complicated non-linear function of experience and educational level) has no significant relationship to parent requests. In contrast, in column 3 we see that teachers who receive higher principal ratings receive more parent requests. In fact, teachers that are rated one standard deviation higher than average receive requests from 4.1 percentage point more parents than the average teacher. Given that the average teacher receives requests from roughly 10 percent of parents in the grade, this reflects a 40 percent increase.[47] Columns 4 and 5 show that the value-added measures of teacher quality are not significantly related to parent requests. These results are mirrored in the R-squared statistics shown in the bottom row. Indeed, bootstrap tests indicate that principal ratings explain

[46] In other work, however, we show that first-ear teachers tend to receive very few requests (Jacob and Lefgren 2005).

[47] For the purpose of the forecast shown in Table X, it is not important that principal ratings may respond to parent requests. In particular, the current specifications show the ability of principal ratings to forecast future requests. Even if principal ratings simply mirror the past preferences of parents, they still are still more useful for predicting subsequent parent satisfaction than are the value-added measures.

significantly more of the variation in parent requests than either the compensation or the value-added measures.

VII. Conclusions

In this paper, we compare subjective principal assessments to the traditional determinants of teacher compensation – education and experience – and to value-added measures of teacher effectiveness based on student achievement gains. We find that subjective principal assessments of teachers predict future student achievement significantly better than teacher experience, education or actual compensation. While value-added measures of teacher effectiveness generally do a better job at predicting future student achievement than principal ratings, the two measures do about equally well in identifying the best and worst teachers. With regard to parent satisfaction, we find that a principal's overall rating of a teacher is a substantially better predictor of future parent requests for that teacher than *either* the teacher's experience, education and current compensation *or* the teacher's value-added achievement measure.

These results suggest that student achievement (as measured by standardized test scores) would likely improve under a principal-based assessment system relative to the current system of compensation. To the extent that the most important staffing decisions involve sanctioning incompetent teachers and/or rewarding the best teachers, a principal-based system may also produce achievement outcomes roughly comparable to a test-based accountability system. In addition, increasing a principal's ability to sanction and reward teachers would likely improve educational outcomes valued by parents but not readily captured by standardized tests.

On the other hand, the inability of principals to distinguish between a broad middle-range of teacher quality suggests caution in relying on principals for fine grained performance

determinations as might be required under certain merit pay policies. Moreover, the evidence that principals may discriminate against male and untenured faculty (and favor certain other teachers) raises some concerns about not only the equity of a principal-based assessment system, but the efficiency as well.

Finally, this paper sheds light on subjective performance measurement more generally. We demonstrate the importance of accounting for estimation error in measured productivity and of looking carefully at the relationship between subjective evaluations and actual productivity across the productivity distribution. Our findings provide further evidence that subjective assessments are only imperfect substitutes for objective measures, and may suffer from important supervisor biases.

References

Aaronson, Daniel, Lisa Barrow and William Sander (2002). "Teachers and student achievement in the Chicago public high schools." Working Paper Series WP-02-28, Federal Reserve Bank of Chicago

Alexander, Elmore R. and Ronnie D. Wilkins (1982). "Performance rating validity: The relationship of objective and subjective measures of performance." Group & Organization Studies 7(4): 485-496.

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 Number R-2007-LAUSD. Santa Monica, CA: RAND Corporation.

Bolino, Mark C. and William H. Turnley (2003). "Counternormative impression management, likeability, and performance ratings: the use of intimidation in an organizational setting." Journal of Organizational Behavior 24(2): 237-250.

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.

Figlio, David and Maurice E. Lucas (2004). "What's in a Grade? School Report Cards and House Prices." American Economic Review. 94(3): 591-604.

Figlio, D., Winicki, J., forthcoming. Food for Thought? The Effects of School Accountability on School Nutrition. Journal of Public Economics.

Hanushek, Eric A. (1997). "Assessing the effects of school resources on student performance: an update." Educational Evaluation and Policy Analysis 19(2): 141-164.

Hanushek, Eric A (1992). "The trade-off between child quantity and quality." Journal of Political Economy 100(1): 84-117.

Hanushek, Eric A. (1986). "The economics of schooling: production and efficiency in public schools." Journal of Economic Literature 49(3): 1141-1177.

Hanushek, Eric A., John Kain, Daniel M. O'Brien and Steven G. Rivkin (2005). "The Market for Teacher Quality." NBER Working Paper #11154.

Hanushek, Eric A. and Steven G. Rivkin (2004). "How to Improve the Supply of High Quality Teachers." In Ravitch, Diane, (ed.), *Brookings Papers on Education Policy 2004*. Washington, DC: Brookings Institution Press.

Heneman, Robert L. (1986). "The relationship between supervisory ratings and results-oriented measures performance: a meta-analysis." Personnel Psychology 39: 811-826.

Heneman, Robert L., David B. Greenberger and Chigozie Anonyuo (1989). "Attributions and exchanges: the effects of interpersonal factors on the diagnosis of employee performance." The Academy of Management Journal 32(2): 466-476.

Jacob, Brian A. (2005). "Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools. Journal of Public Economics. 89(5-6): 761-796.

Jacob, Brian A. and Lars Lefgren (2005). "What Do Parents Value in Education? An Empirical Investigation of Parents' Revealed Preferences for Teachers." Working paper.

Jacob, Brian A. and Stephen D. Levitt (2003). "Rotten apples: an investigation of the prevalence and predictors of teacher cheating." Quarterly Journal of Economics CXVIII(3): 843-878.

Jacobellis v. Ohio. 378 U.S. 184, 197 (1964)

Kane, T., Staiger, D., 2002. Volatility in School Test Scores: Implications for Test-Based Accountability Systems. In: Ravitch, D. (Ed.), Brookings Papers on Education Policy 2002, Brookings Institution, Washington, D.C.

Medley, Donald M. and Homer Coker (1987). "The accuracy of principals' judgments of teacher performance." Journal of Educational Research 80(4): 242-247.

Morris, Carl N. (1983). "Parametric empirical Bayes inference: theory and applications." Journal of the American Statistical Association 78(381): 47-55.

Moulton, Brent R. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." Review of Economics and Statistics, Vol. 72, No. 2, (May, 1990), pp. 334-338.

Murnane, Richard (1975). *The Impact of School Resources on the Learning of Inner-City Children*. Cambrdige, MA: Ballinger Publishing Company.

Prendergast, Candice (1999). "The Provision of Incentives in Firms." Journal of Economic Literature 37(1): 7-63.

Peterson, Kenneth D. (2000). *Teacher Evaluation: A Comprehensive Guide to New Directions and Practices* (2d ed.). Thousand Oaks, CA: Corwin Press.

Peterson, Kenneth D. (1987). "Teacher Evaluation with Multiple and Variable Lines of Evidence." American Educational Research Journal 24(2): 311-317.

Reback, Randall (2005). "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." Working paper.

Rockoff, Jonah E. (2004). "The impact of individual teachers on student achievement: evidence from panel data." American Economic Review 94(2): 247-252.

Sullivan, Daniel G. (2001). "A note on the estimation of regression models with heteroskedastic measurement errors." Working paper 2001-23. Federal Reserve Bank of Chicago.

Todd, Petra and Wolpin, Kenneth (2003). "On the Specification and Estimation of the Production Function for Cognitive Achievement." Economic Journal, February.

Varma, Arup and Linda K. Stroh (2001). "The impact of same-sex LMX dyads on performance evaluations." Human Resource Management 40(4): 309-320.

Wayne, Sandy J. and Gerald R. Ferris (1990). "Influence tactics, affect, and exchange quality in supervisor-subordinate interactions: a laboratory experiment and field study." Journal of Applied Psychology 75(5): 487-499.

Appendix A: Sample Principal Survey Form

We thank you for agreeing to answer the questions in this survey. By answering this survey, you will aid in determining what aspects of teacher effectiveness are most important for students, parents, and principals. Your responses to these surveys will be completely confidential. They will never be revealed to any teachers, administrators, parents, or students. Only statistical averages and correlations will be presented in reports for the district and possible publication in an academic journal.

We will now ask you to rate teachers on the basis of a number of different performance criteria. Please use the following descriptions in rating teachers on a scale of 1 to 10.

1-2: Inadequate – The teacher performs substantially below minimal standards in this area.
3-5: Adequate – The teacher meets minimal standards (but could make substantial improvements in this area).
6-8: Very good – The teacher is highly effective in this area.
9-10: Exceptional – The teacher is among the best I have ever seen in this area (e.g., in the top 1% of teachers).

Part I: Teacher Ratings

Teacher Characteristic	Teacher 1	Teacher 2	Teacher 3	Teacher 4	Teacher 5
Dedication and work ethic					
Organization					
Classroom management					
Raising student math achievement					
Raising student reading achievement					
Role model for students					
Student satisfaction with teacher					
Parent satisfaction with teacher					
Positive relationship with colleagues					
Positive relationship with administrators					
Overall teacher effectiveness					
How many years have you worked with this teacher (in your current school or another school)?					
How many years has this individual been teaching (in your school or another)? Please give your best guess if you are not certain.					

Appendix B: Construction of the Teacher Value-Added Measures

In this section, we describe a number of issues relating to the estimation of the value-added measures used in the analysis. To begin, it is important to note that our math sample consists of a subset of the teachers in our reading sample. With the assistance of district administrators, we conducted detailed interviews with principals to ascertain exactly how students are assigned to classrooms and to explicitly examine how the assignment process may influence our estimates. In many schools, it turns out that students are tracked for math instruction, particularly in sixth grade. Because we do not know a student's math instructor in these cases, we do not construct math value-added measures for this sample. Because reading is never tracked across classrooms (i.e., a child's "homeroom" teacher also instructs the student in reading), we are able to estimate reading value-added measures for all teachers.

Several specification details are worth noting. First, our value-added model differs from standard practice in one respect. To the extent that principals evaluate a teacher relative to other teachers within the school, a value-added indicator that measures effectiveness relative to a district rather than school average will be biased downward.[48] To ensure we identify estimates of teacher quality relative to other teachers *within the same school*, we (a) examine teachers who are in their most recent school (i.e. for the small number of switching teachers, we drop observations from their first school), (b) include school fixed effects and then (c) constrain the teacher fixed effects to sum to zero *within* each school.[49]

Second, because of the way in which most standardized achievement tests are constructed, it is often easier to make improvements at the lower end of the skill distribution than

[48] Typical value added models that simply contain school fixed effects identify teacher quality relative to all teachers (or some omitted teacher) in the district.

[49] The fact that principals are likely using different scales when evaluating teachers makes any correlation between supervisor ratings and a district-wide productivity measure largely uninformative (even in the case where principals were attempting to evaluate their own teachers relative to all others in the district).

at the top end. To the extent that student ability varies across classes, this may bias the teacher fixed effects estimated. We test the sensitivity of our results to this concern in two ways. We estimate models that include third-order polynomials in prior achievement (both reading and math), which allows prior achievement to influence contemporaneous achievement differently depending on a student's pre-test score. We also estimate models where the dependent variable is the one-year achievement gain normalized by student prior ability. Specifically, we divide students into *q* different quantiles based on their prior achievement score, y_{ijkt-1}, and then calculate the mean and standard deviation of achievement *gains* ($g_{ijkt} = y_{ijkt} - y_{ijkt-1}$) for each quantile separately, which we denote as μ_g^q and σ_g^q respectively. We then create standardized gain scores that are mean zero and unit standard deviation *within* each quantile: $G_{ijkt}^q = \left(g_{ijkt}^q - \mu_g^q \right) \Big/ \mu_g^q$. This ensures that each student's achievement gain is compared to the improvement of comparable peers in the value-added models, which should mitigate the potential bias described above.[50] Both strategies yield results comparable to our baseline specification (see Appendix E).

The third specification issue involves the role of teacher experience. Recent evidence indicates that teacher effectiveness at raising achievement increases with experience, particularly in the first few years of a teacher's career (Rockoff 2004, Hanushek et al. 2005). Yet the specification in model (1) implicitly assumes that teacher effectiveness does not vary with experience. To test this assumption, we estimate several variations of model (1) in which we include time-varying indicators of teacher experience along with teacher fixed effects, and calculate each teacher's predicted value-added as of 2002-03, the year in which principals were asked to evaluate the teachers. As expected, new teachers are both substantively and statistically

[50] Hanushek et al. (2005) and Reback (2005) utilize a similar strategy.

less effective than those who have taught for a longer period of time.[51] Despite the importance of experience in determining teacher effectiveness at a point in time, however, Table E1 shows that the correlation between principal ratings and teacher effectiveness is robust to our consideration of teacher experience. This might be because our analysis period is relatively short and our sample includes relatively few novice teachers.

Finally, it is worth noting the inherent limitations of value-added models. As Todd and Wolpin (2003) point out, even if one is not concerned about omitted variables (e.g., when students and teachers are randomly assigned to classes), the $\delta_j's$ will generally not capture the impact of teacher *j* alone, but will also incorporate the effects of optimizing behavior on the part of families. If a child gets randomly assigned to a poor teacher, for example, her parents may spend more time helping the child with schoolwork or enroll her in an afterschool program.[52]

[51] To examine this question, we include a dummy variable that indicates first year teachers. This dummy variable is negative and significant. In specifications in which we include a single log experience measure, we again observe that experienced teachers are more effective one average though the relationship is statistically insignificant. Note that any linear function of experience will be collinear with teacher and year fixed effects. Identification of an experience effect is possible because we include nonlinear measures of experience including a new teacher dummy or log experience.

[52] Moreover, each of the specifications involves implicit assumptions regarding the educational production function. For example, a model that includes lagged achievement measures and contemporaneous school inputs implicitly assumes that the effect of all inputs decay at the same rate. Because we control for lagged achievement, our baseline specification also assumes that the effects of all inputs decay at the same rate but allows for students to progress at different speeds during the year. Our specifications that use a gain score as the outcome assumes that students are on a constant trajectory from the time they enter school (either improving or declining each year) except for the impact of contemporaneous inputs. Furthermore, the effects of transitory changes in educational inputs on the achievement level are assumed to be permanent.

Appendix C: Statistical Properties of Empirical Bayes (EB) Estimates of Teacher Quality

The intuition behind the EB approach is that one can construct more efficient estimates of teacher quality by "shrinking" noisy estimates of teacher effectiveness to the mean of the teacher quality distribution. Because each realization of a teacher's value-added reflects both actual ability as well as measurement error, noisy estimates provide less information regarding true teacher quality than more precise estimates. The EB estimate for teacher *j* is essentially a weighted average of the teacher's fixed effect and the average value-added within the population, where the weight is a function of the reliability of each teacher's fixed effect. This approach is conceptually quite similar to the approach outlined by Sullivan (2001), which involves extending a traditional errors-in-variables correction to allow for heteroskedastic measurement error.

Suppose we have a noisy measure of teacher quality $\hat{\delta}_j = \delta_j + e_j$, where δ_j is actual teacher ability, $\hat{\delta}_j$ is unbiased OLS estimate of teacher ability, and e_j is a mean zero error term. Further assume that both δ_j and e_j are normally distributed with a known mean and variance. If one knew the mean and variance of the distributions of δ_j and e_j, one could construct a more efficient estimate of δ_j that optimally incorporates available intuition. Indeed, it is straightforward to show that $E\left[\delta_j \mid \hat{\delta}_j\right] = \left(1-\lambda_j\right)\bar{\delta} + \left(\lambda_j\right)\hat{\delta}_j$ where $\lambda_j = \frac{\sigma_\delta^2}{\sigma_\delta^2 + \sigma_{e_j}^2}$. The EB estimate for teacher j is exactly this expected value: $E\left[\delta_j \mid \hat{\delta}_j\right] = \hat{\delta}_j^{EB}$. In practice, the mean of the teacher ability distribution, $\bar{\delta}$, is unidentified so all of the effects are centered around zero. Note that we assume that teacher quality is distributed normally with variance σ_δ^2 while $\sigma_{e_j}^2$ is the variance of the measurement error for teacher *j*'s fixed effect, which can vary across observations depending on the amount of data used to construct the estimate.

Of course, the mean of the teacher quality distribution and the variance of the error term are not generally known and must be estimated. One can construct an empirical analog to the expectation above using the method proposed by Morris (1983). This essentially involves using the estimated mean and variance to calculate the appropriate shrinkage factor, λ_j, and incorporating an appropriate degrees of freedom adjustment.[53] We will refer to this estimate as $\hat{\delta}_j^{EB}$. The resulting properties of this EB estimate are essentially the same as if these parameters were known, so for simplicity we will act as if the parameters were known for the remainder of the discussion.

One can easily show that using the EB estimates as an explanatory variable in a regression context will yield point estimates that are unaffected by the attenuation bias that would exist if one used simple OLS estimates. Define the error of the EB estimate as v_j, so that $\delta_j = \hat{\delta}_j^{EB} + v_j$. Because the EB procedure takes advantage of all available information to construct the estimated teacher effect — indeed it is the empirical analog to the conditional expectation of δ_j — the shrinkage estimator is uncorrelated with the error term: $\text{cov}(v, \hat{\delta}^{EB}) = 0$. In fact, the shrinkage estimate can also be thought of as the predicted value from a regression of the actual teacher quality on the noisy measure. By construction, this prediction is orthogonal to the residual v_j.

Too see that the EB estimate of teacher quality will yield unbiased estimates when used as an explanatory variable in a regression context, consider the following simple regression equation:

[53] The degrees of freedom adjustment takes into account that the mean that we are shrinking toward is estimated—not known.

$$
\begin{aligned}
y_j &= \beta_0 + \beta_1 \delta_j + u_j \\
&= \beta_0 + \beta_1 \hat{\delta}_j^{EB} + \beta_1 v_j + u_j
\end{aligned}
\tag{C1}
$$

Because, $\hat{\delta}_j^{EB}$ is orthogonal to the composite error term, $\beta_1 v_j + u_j$, we know the resulting estimate of β_1 will be unbiased.

Though the EB estimates yield unbiased regression coefficients, they do not yield unbiased correlations. Assuming a constant variance for *e*, we can write the variance of the EB estimates as $\sigma^2_{\hat{\delta}^{EB}} = \sigma^2_{\delta}\left(\frac{\sigma^2_{\delta}}{\sigma^2_{\delta} + \sigma^2_{e}}\right) = \sigma^2_{\delta} - \sigma^2_{v}$. In other words, the variance of EB estimates is lower than the variance of actual teacher quality. As the variance of the measurement error in the un-shrunk estimates increases, the variance of EB measures falls and vice versa. As we see below, this implies that the correlations with the EB estimates will be biased downward relative to correlations with the true measure.

$$
\begin{aligned}
corr(y, \hat{\delta}^{EB}) &= \frac{\text{cov}(y, \hat{\delta}^{EB})}{\sigma_y \sqrt{\sigma^2_{\hat{\delta}^{EB}}}} \\
&= \frac{\text{cov}(\beta_0 + \beta_1 \delta + u, \delta - v)}{\sigma_y \sqrt{\sigma^2_{\hat{\delta}^{EB}}}} \\
&= \frac{\beta_1 \sigma^2_{\delta} - \beta_1 \,\text{cov}(\delta, v)}{\sigma_y \sqrt{\sigma^2_{\hat{\delta}^{EB}}}} \\
&= \frac{\beta_1 (\sigma^2_{\delta} - \sigma^2_{v})}{\sigma_y \sqrt{\sigma^2_{\delta} - \sigma^2_{v}}} \\
&= \frac{\beta_1 \sigma_{\delta} \sqrt{\sigma^2_{\delta} - \sigma^2_{v}}}{\sigma_{\delta} \sigma_y} < \frac{\beta_1 \sigma^2_{\delta}}{\sigma_{\delta} \sigma_y} = corr(y, \delta)
\end{aligned}
\tag{C2}
$$

Now suppose that it is known that the distribution of value added varies across a set of *K* different groups. For example, the distribution of actual teacher quality may vary by gender or experience. In this case, the conditional expectation of δ_j is

$E\left(\delta_j \mid \hat{\delta}_j, group = k\right) = \left(1 - \lambda_j\right)\overline{\delta}_k + \lambda_j \hat{\delta}_j$, where $\overline{\delta}_k$ is the mean of teacher quality of teachers in group *k*. Additionally, λ_j must be constructed using the variance of δ_j around the group-specific mean. Morris' (1983) method of constructing EB estimates readily generalizes to this situation, though in practice it may be necessary to impose substantial structure on the conditional mean.[54] The advantage of allowing the mean of the teacher quality measure to vary with covariates is that one can generate more precise estimates of teacher quality. Furthermore, the error of the EB estimate will be orthogonal to every piece of information (e.g. gender) used to construct it. This guarantees regression coefficient estimates that are unbiased by measurement error in a context that includes covariates besides the EB measure itself.

[54] For example, one may need to assume that the conditional mean of teacher ability is a quadratic function of experience to conserve degrees of freedom.

Appendix D:

Non-Parametric Measures of Association between Performance Indicators

In order to get a more intuitive understanding of the magnitude of the relationship between principal ratings and actual teacher effectiveness, we calculate several simple, non-parametric measures of the association between the subjective and objective performance indicators. While this exercise is complicated somewhat by the existence of measurement error in the teacher value-added estimates, it is relatively straightforward to construct such measures through Monte Carlo simulations with only minimal assumptions. Following the notation in the text, we define the principal's assessment of teacher *j* as $\hat{\delta}_j^P$, the estimated value-added of teacher *j* as $\hat{\delta}_j$ and the true ability of teacher *j* as δ_j. Our goal is to calculate the following probabilities:

(D1) $\Pr\left(\delta_j = t \mid \hat{\delta}_j^P = t\right)$

(D2) $\Pr\left(\delta_j = b \mid \hat{\delta}_j^P = b\right)$

where *t* (*b*) indicates that the teacher was in the top (bottom) quantile of the distribution. For example, (D1) is the probability that the teacher is in the top quantile of the true ability distribution conditional on being in the top quantile of the distribution according to the principal assessment.

We can calculate the conditional probability of a teacher's value-added ranking given her principal ranking directly from the data. These probabilities can be written as follows:

(D3)

$$\Pr\left(\hat{\delta}_j = t \mid \hat{\delta}_j^P = t\right) = \Pr\left(\hat{\delta}_j = t \mid \delta_j = t\right)\Pr\left(\delta_j = t \mid \hat{\delta}_j^P = t\right) + \Pr\left(\hat{\delta}_j = t \mid \delta_j = b\right)\Pr\left(\delta_j = b \mid \hat{\delta}_j^P = t\right)$$

(D4)

$$\Pr\left(\hat{\delta}_j = b \mid \hat{\delta}_j^P = t\right) = \Pr\left(\hat{\delta}_j = b \mid \delta_j = t\right)\Pr\left(\delta_j = t \mid \hat{\delta}_j^P = t\right) + \Pr\left(\hat{\delta}_j = b \mid \delta_j = b\right)\Pr\left(\delta_j = b \mid \hat{\delta}_j^P = t\right)$$

(D5)

$$\Pr\left(\hat{\delta}_j = t \mid \hat{\delta}_j^P = b\right) = \Pr\left(\hat{\delta}_j = t \mid \delta_j = t\right)\Pr\left(\delta_j = t \mid \hat{\delta}_j^P = b\right) + \Pr\left(\hat{\delta}_j = t \mid \delta_j = b\right)\Pr\left(\delta_j = b \mid \hat{\delta}_j^P = b\right)$$

(D6)

$$\Pr\left(\hat{\delta}_j = b \mid \hat{\delta}_j^P = b\right) = \Pr\left(\hat{\delta}_j = b \mid \delta_j = t\right)\Pr\left(\delta_j = t \mid \hat{\delta}_j^P = b\right) + \Pr\left(\hat{\delta}_j = b \mid \delta_j = b\right)\Pr\left(\delta_j = b \mid \hat{\delta}_j^P = b\right)$$

Note that the four equations also assume that the fact that the principal rates a teacher in the top (bottom) category does not provide any additional information regarding whether the OLS measure of the value-added will be in the top (bottom) category once we condition on whether the teacher's true ability is in the top (bottom) category. For example, in equation (D3), we assume that $\Pr\left(\hat{\delta}_j = t \mid \delta_j = t\right) = \Pr\left(\hat{\delta}_j = t \mid \delta_j = t, \hat{\delta}_j^P = t\right)$ $\Pr\left(\hat{\delta}_j = t \mid \delta_j = b\right) = \Pr\left(\hat{\delta}_j = t \mid \delta_j = b, \hat{\delta}_j^P = t\right)$. While we do not believe this is strictly true, it should not substantially bias our estimates.

We also know the following identities are true:

(D7) $\Pr\left(\delta_j = t \mid \hat{\delta}_j^P = t\right) + \Pr\left(\delta_j = b \mid \hat{\delta}_j^P = t\right) = 1$

(D8) $\Pr\left(\delta_j = b \mid \hat{\delta}_j^P = b\right) + \Pr\left(\delta_j = t \mid \hat{\delta}_j^P = b\right) = 1$

(D9) $\Pr\left(\hat{\delta}_j = t \mid \hat{\delta}_j^P = t\right) + \Pr\left(\hat{\delta}_j = b \mid \hat{\delta}_j^P = t\right) = 1$

(D10) $\Pr\left(\hat{\delta}_j = t \mid \hat{\delta}_j^P = b\right) + \Pr\left(\hat{\delta}_j = b \mid \hat{\delta}_j^P = b\right) = 1$

We can solve (D3) and (D7) to obtain (D1) as follows:

$$\text{(D11)}\quad \begin{aligned}\Pr\left(\delta_j = t \mid \hat{\delta}_j^P = t\right) &= 1 - \left[\frac{\Pr\left(\hat{\delta}_j = t \mid \hat{\delta}_j^P = t\right) - \Pr\left(\hat{\delta}_j = t \mid \delta_j = t\right)}{\Pr\left(\hat{\delta}_j = t \mid \delta_j = b\right) - \Pr\left(\hat{\delta}_j = t \mid \delta_j = t\right)}\right] \\ &= \frac{\Pr\left(\hat{\delta}_j = t \mid \hat{\delta}_j^P = t\right) - \Pr\left(\hat{\delta}_j = t \mid \delta_j = b\right)}{\Pr\left(\hat{\delta}_j = t \mid \delta_j = t\right) - \Pr\left(\hat{\delta}_j = t \mid \delta_j = b\right)}\end{aligned}$$

Using Bayes' Rule, we can rewrite (D11) as follows:

$$\text{(D12)}\quad \Pr\left(\delta_j = t \mid \hat{\delta}_j^P = t\right) = \frac{\Pr\left(\hat{\delta}_j = t \mid \hat{\delta}_j^P = t\right) - \Pr\left(\delta_j = b \mid \hat{\delta}_j = t\right)\dfrac{\Pr\left(\hat{\delta}_j = t\right)}{\Pr\left(\delta_j = b\right)}}{\Pr\left(\delta_j = t \mid \hat{\delta}_j = t\right)\dfrac{\Pr\left(\hat{\delta}_j = t\right)}{\Pr\left(\delta_j = t\right)} - \Pr\left(\delta_j = b \mid \hat{\delta}_j = t\right)\dfrac{\Pr\left(\hat{\delta}_j = t\right)}{\Pr\left(\delta_j = b\right)}}$$

We can estimate all of the remaining quantities in (D12) from our data. More specifically, we can calculate estimates of the following probabilities through simulation:

$$\text{(D13)}\quad \Pr\left(\delta_j = t \mid \hat{\delta}_j = t\right)$$

$$\text{(D14)}\quad \Pr\left(\delta_j = b \mid \hat{\delta}_j = t\right)$$

$$\text{(D15)}\quad \Pr\left(\delta_j = t \mid \hat{\delta}_j = b\right)$$

$$\text{(D16)}\quad \Pr\left(\delta_j = b \mid \hat{\delta}_j = b\right)$$

To do so, we assume that the true ability of teacher *j* is distribution normally with a mean equal to the estimated value-added for teacher $j, \hat{\delta}_j$, and a variance equal to $Var\left(\hat{\delta}_j\right)$. We then randomly draw 500 realizations of each teacher's true ability, $\hat{\delta}_j$, and for each draw determine which set of teachers would fall in the top (bottom) quantile of the ability distribution and whether the principal would have correctly classified the teacher based on this realization. We estimate the probabilities in (D13) – (D16) as the average of these realizations. Finally, we can

calculate $\Pr\left(\hat{\delta}_j = t\right) = \Pr\left(\delta_j = t\right)$ and $\Pr\left(\hat{\delta}_j = b\right) = \Pr\left(\delta_j = b\right)$ directly from our original data. In many cases, for example, because we are interested in the top versus bottom quantiles, we know that $\Pr\left(\hat{\delta}_j = t\right) = \Pr\left(\delta_j = t\right) = \Pr\left(\hat{\delta}_j = b\right) = \Pr\left(\delta_j = b\right)$, so that the ratios in (D12) will cancel out. For example, the proportion of teachers in the top half of the true ability distribution will be 0.50 by definition, as will be the proportion of teachers in the top half of the value-added distribution.

In a similar fashion, we can obtain (D2) by solving (D5) and (D8):

$$\Pr\left(\delta_j = b \mid \hat{\delta}_j^P = b\right) = \frac{\Pr\left(\hat{\delta}_j = b \mid \hat{\delta}_j^P = b\right) - \Pr\left(\hat{\delta}_j = b \mid \delta_j = t\right)}{\Pr\left(\hat{\delta}_j = b \mid \delta_j = b\right) - \Pr\left(\hat{\delta}_j = b \mid \delta_j = t\right)}$$

(D17)

$$= \frac{\Pr\left(\hat{\delta}_j = b \mid \hat{\delta}_j^P = b\right) - \Pr\left(\delta_j = t \mid \hat{\delta}_j = b\right)\dfrac{\Pr\left(\hat{\delta}_j = b\right)}{\Pr\left(\delta_j = t\right)}}{\Pr\left(\delta_j = b \mid \hat{\delta}_j = b\right)\dfrac{\Pr\left(\hat{\delta}_j = b\right)}{\Pr\left(\delta_j = b\right)} - \Pr\left(\delta_j = t \mid \hat{\delta}_j = b\right)\dfrac{\Pr\left(\hat{\delta}_j = b\right)}{\Pr\left(\delta_j = t\right)}}$$

Appendix E: Sensitivity Analyses

Table E1 presents a series of sensitivity analyses for the main results regarding the relationship between principal ratings and teacher value-added. In short, the results presented above are robust to a variety of alternative specifications. Row 1 displays the baseline results discussed earlier. Row 2 shows the reading results using the sample of teachers for whom math value-added measures are also available. Rows 3-9 show that that the estimated correlations are robust to using a variety of different value-added specifications. In particular, rows 3-5 show results that use value-added measures derived from models that include student fixed effects or control for student prior achievement differently. Rows 6-7 use value-added specifications that account for the test measurement issues described in Section IV. Rows 8-9 use a value-added measure of each teacher's predicted "quality" as of 2002-03 (the time of the principal evaluations) that is constructed from models that incorporate time-varying measures of teacher experience as well as teacher fixed effects.

Rows 10-12 show that the estimated correlations are not sensitive to excluding particular groups such as first-year teachers, schools with lumpy principal ratings or teachers that did not have test score data in 1998-2002. The one notable exception, which was discussed earlier and is shown in Figure III, involves excluding teachers who received the top or bottom principal ratings (row 12). If these teachers are excluded from the sample, the correlations are very small in magnitude and not significantly different than zero, indicating that principals have no ability to distinguish between teachers in the middle of the distribution.

Row 14 presents correlations that are based on the value-added measures derived from 1998-2002 test score data. Because principals were asked to evaluate teachers in the middle of the 2002-03 school year, it is possible that they would not have had an opportunity to fully

observe teacher performance in this year, which might bias downward any results that incorporate the data from this year. However, the results in row 14 are nearly identical to those in row 12 (which includes the same sample of teachers but use the value-added measures based on 1998-2003 data), suggesting that this is not a serious concern in practice. Row 14 shows the correlation between the teacher value-added and the principal's overall rating of the teacher. The reading results are nearly identical to baseline, though the correlations for math are somewhat smaller (though not significantly so) as one would expect.

It is possible that principals value the improvement of students at different points on the ability distribution differently. For example, principals may be most concerned with the lowest achieving students in their school. If this were the case, principals may evaluate teachers on the basis of their performance with these students rather than with all students in the teacher's class. The estimates in rows 15-19 examine this possibility. Row 15 shows correlations between principal ratings and a value-added measure that uses a binary proficiency indicator instead of a continuous test score as the outcome measure.[55] The results are roughly comparable to the baseline.

Rows 16-19 show estimated correlations that use value-added measures which are based on student performance of either high or low achieving students. For example, the value-added measures used in row 16 include only students whose prior achievement score was below the district average. Row 18 includes only students in the bottom half of his or her classroom distribution that year. In all cases, the outcome measure is a normalized gain score which takes

[55] The Core exams are criterion-referenced and student results are reported in terms of four different proficiency levels: minimal mastery, partial mastery, near mastery, mastery. Discussions with district officials suggest that principals and teachers focused primarily on whether children reached level three, near mastery, because students scoring at level one or two were typically considered candidates for remedial services. For this reason, we define proficient as scoring at level 3 or 4. Our results are robust to alternative classifications. The results are also comparable when we use a Logit or Probit model instead of OLS.

into account the achievement quantile in which the student started, thus allowing one to accurately compare student performance across different ranges of the ability distribution.

The results are quite interesting. The adjusted correlations that use value-added measures based on the bottom half of the achievement distribution are roughly twice as large as those correlations that use value-added measures based on above average students. While these estimates are not precise enough to distinguish statistically, these results suggest that principals may indeed focus disproportionately on how teachers perform with the lowest achieving students in their classes.

Since it is possible that principals may be more aware of the ability of certain teachers than others, it is interesting to examine the correlation between principal assessment and value-added for various subgroups of teachers. Unfortunately, our sample size limits the ability to draw strong conclusions from such comparisons, particularly in the cases of teacher experience. The estimates in rows 24-25, however, suggest that principals may be better at identifying teacher effectiveness in math for teachers in the younger grades relative to the older grades.

It is also interesting to consider whether principals are aware of their ability (or lack thereof) at recognizing which teachers are effective. As part of the survey, we asked principals to judge how confident they felt in each of their ratings. Principals who indicate that they are "very" or "completely" confident gave ratings that were significantly and substantially more correlated with teacher productivity than their peers. In rows 26-29, for example, we see that principals who are confident in their reading or math ratings have correlations that are two to four times larger than their less confident colleagues. This suggests that there may be considerably heterogeneity in principal ability which might be explored in subsequent research.

Finally, the comparison of male vs. female principals in rows 30-31 provides some evidence that male principals may be better ability to identify effective math teachers, although one would not want to place a causal interpretation on this result since it is possible that principal background/training varies by gender in ways that influence familiarity with subject matter.

TABLE E1
SENSITIVITY ANALYSES

	Specification	Correlation between value-added measure and principal rating of teacher's ability to raise math (reading) achievement			
		Reading		Math	
		Raw	Adj.	Raw	Adj.
1	Baseline	0.20* (0.06)	0.32* (0.10)	0.28* (0.07)	0.36* (0.09)
2	For math sample	0.26* (0.08)	0.36* (0.10)	--	--
3	Outcome is gain score	0.20* (0.06)	0.30* (0.09)	0.28* (0.07)	0.37* (0.09)
4	Level outcome with student fixed effects	0.19* (0.07)	0.28* (0.12)	0.28* (0.09)	0.39* (0.12)
5	Gain outcome with student fixed effects	0.19* (0.06)	0.35† (0.19)	0.33* (0.07)	0.47* (0.10)
6	Outcome is the gain score normalized around predicted gain for students with comparable prior achievement	0.21* (0.07)	0.32* (0.10)	0.29* (0.07)	0.37* (0.09)
7	Outcome is achievement level but polynomials in prior achievement are included as covariates	0.20* (0.06)	0.32* (0.10)	0.28* (0.07)	0.35* (0.09)
8	Include indicator for first-year teachers (+ polynomials in prior achievement)	0.20* (0.07)	0.31* (0.10)	0.28* (0.07)	0.36* (0.09)
9	Include ln(experience) variable (+ polynomials in prior achievement)	0.20* (0.06)	0.32* (0.09)	0.29* (0.07)	0.38* (0.10)
10	Exclude schools with extreme lumpiness in principal ratings	0.17* (0.07)	0.30* (0.13)	0.28* (0.08)	0.37* (0.11)
11	Exclude teachers who were in their first year in 2002-03	0.21* (0.07)	0.29* (0.09)	0.29* (0.07)	0.36* (0.09)
12	Exclude teachers with top or bottom principal rating	0.03 (0.07)	0.06 (0.14)	-0.01 (0.09)	-0.02 (0.11)
13	Use value-added measures created from 1998-2002 test score data (which automatically excludes teachers who only had test score data in 2002-03, including first-year teachers: sample is identical to row 12)	0.19* (0.07)	0.25* (0.09)	0.36* (0.08)	0.46* (0.09)
14	Use principal's overall rating of the teacher (instead of the principal's rating of teacher's ability to raise student achievement)	0.20* (0.07)	0.31* (0.11)	0.19* (0.08)	0.24* (0.10)
15	Use proficiency measure as outcome	0.11 (0.07)	0.18 (0.12)	0.13† (0.08)	0.19 (0.12)
16	Estimate VA using only students <u>below district</u> mean achievement – use normalized gain score as outcome	0.23* (0.06)	0.41* (0.13)	0.29* (0.07)	0.46* (0.13)
17	Estimate VA using only students <u>above district</u> mean achievement – use normalized gain score as outcome	0.12† (0.07)	0.22 (0.16)	0.27* (0.08)	0.35* (0.10)
18	Estimate VA using only students <u>below class</u> mean achievement – use normalized gain score as outcome	0.22* (0.06)	0.38* (0.11)	0.28* (0.07)	0.41* (0.10)
19	Estimate VA using only students <u>above class</u> mean achievement – use normalized gain score as outcome	0.11 (0.07)	0.18 (0.13)	0.26* (0.08)	0.33* (0.10)

	By teacher characteristics				
20	Experienced teachers (>=11 years, n=94)	0.32*	0.39*	0.36*	0.42*
		(0.08)	(0.10)	(0.08)	(0.09)
21	Inexperienced teachers (<11 years, n=108)	0.08	0.74	0.16	0.25
		(0.08)	(1.97)	(0.11)	(0.20)
22	Principal known teacher for long time (>=4 years, n=114)	0.25*	0.32*	0.31*	0.37*
		(0.09)	(0.10)	(0.08)	(0.09)
23	Principal hasn't known teacher for long (<4 years, n=88)	0.14	0.46	0.25*	0.34
		(0.09)	(0.47)	(0.11)	(0.19)
24	Grades 2-4 (n=127)	0.31*	0.44*	0.38*	0.48*
		(0.07)	(0.10)	(0.07)	(0.09)
25	Grades 5-6 (n=75)	0.13	0.64	-0.04	-0.06
		(0.10)	(0.88)	(0.14)	(0.26)
	By principal characteristics				
26	Principal confident in reading ratings (n=107)	0.23*	0.39*	0.39*	0.51*
		(0.08)	(0.15)	(0.09)	(0.11)
27	Principal not confident in reading ratings (n=79)	0.13	0.19	0.15	0.19
		(0.11)	(0.16)	(0.12)	(0.15)
28	Principal confident in math ratings (n=48)	0.35*	0.60	0.55*	0.87†
		(0.10)	(0.83)	(0.09)	(0.62)
29	Principal not confident in math ratings (n=138)	0.13†	0.19	0.20*	0.24*
		(0.08)	(0.12)	(0.09)	(0.12)
30	Male principal (n=108)	0.25*	0.38*	0.39*	0.53*
		(0.08)	(0.14)	(0.09)	(0.11)
31	Female principal (n=94)	0.13	0.19	0.20*	0.24*
		(0.08)	(0.12)	(0.10)	(0.13)

Notes: The adjusted correlations take into account the estimation error in our value-added measures of teacher effectiveness. Bootstrapped standard errors are in parentheses.
* = significant at the 5 percent level; † = significant at the 10 percent level.

TABLE I
SUMMARY STATISTICS

Student Characteristics	Mean
Male	0.51
White	0.73
Black	0.01
Hispanic	0.21
Other	0.06
Limited English Proficiency	0.21
Free or Reduced Price Lunch	0.48
Special Education	0.12
Math Achievement (national percentile)	0.49
Reading Achievement (national percentile)	0.49
Language Achievement (national percentile)	0.47
Teacher Characteristics	Mean (s.d.)
Male	0.16
Age	41.9 (12.5)
Untenured	0.17
Experience	11.9 (8.9)
Fraction with 10-15 Years Experience	0.19
Fraction with 16-20 Years Experience	0.14
Fraction with 21+ Years Experience	0.16
Years working with principal	4.8 (3.6)
BA Degree at in state (but not local) college	0.10
BA Degree at out of state college	0.06
MA Degree	0.16
Any additional endorsements	0.20
Any additional endorsements in areas other than ESL	0.10
Licensed in more than one area	0.27
Licensed in area other than ECE or EE	0.07
2nd Grade	0.24
3rd Grade	0.21
4th Grade	0.20
5th Grade	0.18
6th Grade	0.18
Mixed grade classroom	0.08
Two teachers in the classroom	0.05
Number of teachers	202
Number of principals	13

Notes: Student characteristics are based on students enrolled in grades 2-6 in Spring 2003. Math and reading achievement measures are based on the Spring 2002 scores on the Stanford Achievement Test (Version 9) taken by selected elementary grades in the district. Teacher characteristics are based on administrative data. Nearly all teachers in the district are Caucasian, so race indicators are omitted.

TABLE II
SUMMARY STATISTICS FOR PRINCIPAL RATINGS

Item	Mean (s.d.)	10^{th} Percentile	90^{th} Percentile
Overall teacher effectiveness	8.07 (1.36)	6.5	10
Dedication and work ethic	8.46 (1.54)	6	10
Organization	8.04 (1.60)	6	10
Classroom management	8.06 (1.63)	6	10
Raising student math achievement	7.89 (1.30)	6	9
Raising student reading achievement	7.90 (1.44)	6	10
Role model for students	8.35 (1.34)	7	10
Student satisfaction with teacher	8.36 (1.20)	7	10
Parent satisfaction with teacher	8.28 (1.30)	7	10
Positive relationship with colleagues	7.94 (1.72)	6	10
Positive relationship with administrators	8.30 (1.66)	6	10

Notes: These statistics are based on the 202 teachers included in the analysis sample.

TABLE III
FACTOR LOADINGS DERIVED FROM PRINCIPAL RATING ITEMS

Item	Factor 1 - Student Satisfaction	Factor 2 – Achievement	Factor 3 - Collegiality
Dedication and work ethic	-0.143	0.671	0.184
Organization	0.017	0.819	0.014
Classroom management	0.226	0.805	-0.197
Raising student math achievement	-0.001	0.767	0.037
Raising student reading achievement	-0.001	0.779	0.051
Role model for students	0.528	0.187	0.245
Student satisfaction with teacher	0.944	-0.037	0.053
Parent satisfaction with teacher	0.677	0.180	0.073
Positive relationship with colleagues	0.101	-0.032	0.812
Positive relationship with administrators	0.066	-0.030	0.853

Notes: Factors derived from ML factor analysis with a Promax rotation. All individual survey items are normalized to have a mean of zero and standard deviation of one within schools.

TABLE IV
CORRELATION BETWEEN A PRINCIPAL'S RATING OF A TEACHER'S ABILITY TO RAISE STUDENT ACHIEVEMENT AND THE VALUE-ADDED MEASURE OF THE TEACHER'S EFFECTIVENESS AT RAISING STUDENT ACHIEVEMENT

	Reading		Math		
	Unadjusted	Adjusted	Unadjusted	Adjusted	Diff: Reading – Math (2) – (4)
	(1)	(2)	(3)	(4)	(5)
(1) Using baseline specification for creating value-added measure	0.20* (0.07)	0.32* (0.10)	0.28* (0.07)	0.36* (0.09)	-0.00 (0.10)
(2) Using average student achievement (levels) as the value-added measure	0.35* (0.05)	0.56* (0.09)	0.29* (0.08)	0.38* (0.11)	0.20 (0.13)
(3) Difference: (1) – (2)		-0.24* (0.09)		0.02 (0.07)	

Notes: Number of observations for reading and math is 202 and 151 respectively. Adjusted correlations are described in the text. The standard errors shown in parentheses are calculated using a bootstrap. The reading - math difference in column 5 does not equal the simple difference between the values in columns 2 and 4 because the difference is calculated using the limited sample of teachers for whom math value-added measures are available.
* = significant at the 5 percent level; † = significant at the 10 percent level.

TABLE V
RELATIONSHIP BETWEEN PRINCIPAL RATINGS OF A TEACHER'S ABILITY TO RAISE STUDENT ACHIEVEMENT AND TEACHER VALUE-ADDED

	Reading	Math
Conditional probability that a teacher who received the **top rating** from the principal was the top teacher according to the value-added measure (standard error)	0.52 (0.16)	0.69 (0.13)
Null hypothesis (probability expected if principals randomly assigned teacher ratings)	0.14	0.26
Z-score (p-value) of test of difference between observed and null	2.32 (0.02)	3.34 (0.00)
Conditional probability that a teacher who received a rating **above the median** from the principal was above the median according to the value-added measure (standard error)	0.49 (0.10)	0.61 (0.13)
Null hypothesis (probability expected if principals randomly assigned teacher ratings)	0.33	0.24
Z-score (p-value) of test of difference between observed and null	1.58 (0.11)	2.72 (0.01)
Conditional probability that a teacher who received a rating **below the median** from the principal was below the median according to the value-added measure (standard error)	0.49 (0.09)	0.54 (0.12)
Null hypothesis (probability expected if principals randomly assigned teacher ratings)	0.36	0.26
Z-score (p-value) of test of difference between observed and null	1.49 (0.14)	2.30 (0.02)
Conditional probability that the teacher(s) who received the bottom rating from the principal was the **bottom teacher(s)** according to the value-added measure (standard error)	0.42 (0.19)	0.69 (0.13)
Null hypothesis (probability expected if principals randomly assigned teacher ratings)	0.09	0.23
Z-score (p-value) of test of difference between observed and null	1.71 (0.09)	3.51 (0.00)

Notes: The probabilities are calculated using the procedure described in Appendix C.

TABLE VI
PREDICTORS OF OVERALL PRINCIPAL RATING

	Dependent Variable = Principal's Overall Rating of the Teacher							
Independent Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student achievement factor	0.56*		0.56*					0.54*
	(0.04)		(0.04)					(0.05)
Collegiality factor	0.35*		0.37*					0.37*
	(0.04)		(0.04)					(0.05)
Student satisfaction factor	0.15*		0.13*					0.13*
	(0.04)		(0.05)					(0.05)
Number of parent requests		4.25*	0.75				3.44*	0.70
		(1.00)	(0.48)				(0.98)	(0.50)
Reading value-added (EB measure)				1.92*			1.63	-0.02
				(0.64)			(0.90)	(0.44)
Math value-added (EB measure)					0.89*		0.22	0.07
					(0.36)		(0.36)	(0.17)
Male						-0.16	-0.02	-0.07
						(0.19)	(0.20)	(0.09)
Untenured						-0.48*	-0.36	-0.03
						(0.22)	(0.23)	(0.11)
Years of Experience						-0.02*	-0.02†	-0.00
						(0.01)	(0.01)	(0.00)
Grade 2-4						0.14	0.44*	0.01
						(0.15)	(0.19)	(0.09)
Years known principal						0.00	-0.01	0.00
						(0.02)	(0.02)	(0.01)
BA Degree at in state college						0.70*	0.54*	0.13
						(0.23)	(0.22)	(0.10)
BA Degree at out of state college						0.18	0.20	0.11
						(0.29)	(0.29)	(0.13)
MA Degree						0.04	-0.13	0.05
						(0.19)	(0.19)	(0.09)
Additional endorsements						0.15	0.14	0.06
						(0.18)	(0.17)	(0.08)
R-Squared	0.837	0.100	0.840	0.042	0.045	0.141	0.248	0.847

Notes: N=201. Observations with missing teacher background data are set to zero and missing data indicators are included. All models also include school fixed effects. Standard errors are included in parenthesis. * = significant at the 5 percent level; † = significant at the 10 percent level.

TABLE VII
DO PRINCIPALS DISCRIMINATE?

Independent Variables	Dependent Variable= Principal Rating of Teacher Ability to Raise Achievement					
	Reading			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
EB estimate of teacher effectiveness	--	2.34* (0.73)	1.90* (0.67)	--	1.19* (0.34)	1.13* (0.32)
Male	-0.54* (0.21)	-0.33 (0.21)	-0.41* (0.19)	-0.48† (0.26)	-0.50* (0.24)	-0.62* (0.22)
Untenured	-0.62* (0.25)	-0.55* (0.25)	-0.48* (0.21)	-0.64* (0.30)	-0.35 (0.25)	-0.39† (0.22)
Years of Experience	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
Grade 2-4	0.19 (0.16)	0.45* (0.17)	0.34* (0.16)	0.03 (0.21)	0.11 (0.19)	0.00 (0.18)
Years known principal	-0.02 (0.03)	-0.03 (0.03)	-0.01 (0.02)	-0.06 (0.05)	-0.01 (0.02)	-0.01 (0.02)
BA Degree at in state (but not local) college	0.30 (0.25)	0.19 (0.24)	0.01 (0.21)	0.53* (0.30)	0.46† (0.26)	0.28 (0.27)
BA Degree at out of state college	0.33 (0.33)	0.31 (0.32)	0.22 (0.27)	0.10 (0.38)	-0.02 (0.33)	-0.14 (0.31)
MA Degree	-0.05 (0.20)	-0.25 (0.20)	-0.18 (0.18)	0.40† (0.22)	0.25 (0.20)	0.24 (0.19)
Additional endorsements	0.18 (0.19)	0.13 (0.19)	0.14 (0.16)	0.13 (0.22)	0.09 (0.19)	0.09 (0.18)
Relationship with administration			0.32* (0.07)			0.33* (0.07)
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
R-Squared	0.14	0.18	0.29	0.10	0.19	0.31
Observations	202	202	202	151	151	151

Notes: Observations with missing teacher background data are set to zero and missing data indicators are included. All models also include school fixed effects. The EB estimates incorporate information on observed value-added as well as all of the other variables shown in this table. Standard errors are included in parenthesis.
* = significant at the 5 percent level; † = significant at the 10 percent level.

TABLE VIII
ARE CERTAIN TEACHERS MORE EFFECTIVE THAN OTHERS?

	Dependent Variable: Value-Added Measure of Teacher Effectiveness	
Independent Variables	Reading	Math
Male	-0.07†	0.00
	(0.04)	(0.09)
Untenured	-0.04	-0.08
	(0.05)	(0.09)
Years of Experience	0.00	-0.00
	(0.00)	(0.01)
Grade 2-4	-0.10*	-0.08
	(0.03)	(0.07)
Years known principal	0.00	0.01
	(0.00)	(0.01)
BA Degree at in state (but not local) college	0.02	0.04
	(0.05)	(0.10)
BA Degree at out of state college	-0.03	0.09
	(0.06)	(0.12)
MA Degree	0.09*	0.09
	(0.04)	(0.07)
Additional endorsements	0.03	0.02
	(0.03)	(0.07)
Relationship with administration	0.03†	0.01
	(0.01)	(0.03)
School Fixed Effects	Yes	Yes
R-Squared	0.11	0.05
Observations	202	151

Notes: Observations with missing teacher background data are set to zero and missing data indicators are included. All models also include school fixed effects. Standard errors are included in parenthesis.
* = significant at the 5 percent level; † = significant at the 10 percent level.

TABLE IX
THE ASSOCIATION BETWEEN DIFFERENT TEACHER QUALITY MEASURES AND FUTURE STUDENT ACHIEVEMENT

	Dependent Variable											
	2003 Reading Score						2003 Math Score					
Independent Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
3-5 years of experience	-0.023						0.207					
	(0.065)						(0.137)					
6-10 years of experience	0.014						0.018					
	(0.065)						(0.136)					
11-20 years of experience	-0.031						0.113					
	(0.058)						(0.130)					
21+ years of experience	-0.063						-0.003					
	(0.076)						(0.142)					
MA Degree	0.097*						0.088					
	(0.044)						(0.074)					
Annual pay (in $1,000)		-0.000				-0.002		-0.000				-0.000
		(0.002)				(0.003)		(0.004)				(0.003)
Overall principal rating			0.058*			0.036†			0.137*			0.074*
			(0.020)			(0.019)			(0.023)			(0.023)
Principal rating of ability to raise reading (math) scores				0.048*						0.102*		
				(0.020)						(0.030)		
Reading (math) value-added (EB measure)					0.094*	0.085*					0.211*	0.181*
					(0.016)	(0.017)					(0.023)	(0.024)
R-squared	0.476	0.474	0.477	0.476	0.482	0.483	0.387	0.381	0.396	0.389	0.411	0.415

Notes: Each column represents a separate specification. Specifications in columns 1-6 include 160 teachers and 3,834 students; columns 7-12 include 116 teachers and 2,566 students. All regressions include the following variables: male, special education status, free lunch eligibility, limited English proficiency, age, minority, fixed effects for grade and school, lagged math and reading score, class size, class-level average of student demographics and lagged achievement scores, and an indicator for a mixed grade class. Standard errors clustered at the teacher (i.e., classroom) level are shown in parenthesis.
* = significant at the 5 percent level; † = significant at the 10 percent level.

TABLE X
THE ASSOCIATION BETWEEN DIFFERENT TEACHER QUALITY MEASURES AND FUTURE PARENT REQUESTS

	Dependent Variable = Normalized Number of Parent Requests						
Independent Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
3-5 years of experience	-0.032						
	(0.030)						
6-10 years of experience	-0.061*						
	(0.027)						
11-20 years of experience	-0.072*						
	(0.030)						
21+ years of experience	-0.088*						
	(0.034)						
MA Degree	0.025						
	(0.027)						
Annual pay (in $1,000)		-0.001				0.000	-0.001
		(0.002)				(0.002)	(0.002)
Overall principal rating			0.041*			0.043*	0.037*
			(0.012)			(0.013)	(0.014)
Reading value-added (EB measure)				0.002		-0.010	
				(0.012)		(0.011)	
Math value-added (EB measure)					0.007		-0.011
					(0.016)		(0.014)
R-squared	0.128	0.076	0.246	0.067	0.222	0.253	0.385

Notes: Sample includes grades 2-6. The unit of observation is the teacher-grade-school-year. For specifications 1-4 and 6, we have 213 observations representing 144 different teachers. For specifications 5 and 7 we have 156 observations representing 107 different teachers. The dependent variable is the number of parent requests, normalized by subtracting off the average number of requests in the grade-school-year and then dividing by the student enrollment in the grade-school-year. All models include fixed effects for school-grade-year. The fraction of parents requesting any teacher in the sample is 0.283 and the average fraction of parents requesting any individual teacher is 0.095. Standard errors are clustered by teacher. * = significant at 5 percent level. † = significant at 10 percent level.

FIGURE I
THE DISTRIBUTION OF PRINCIPAL RATINGS OF A
TEACHER'S ABILITY TO RAISE STUDENT ACHIEVEMENT

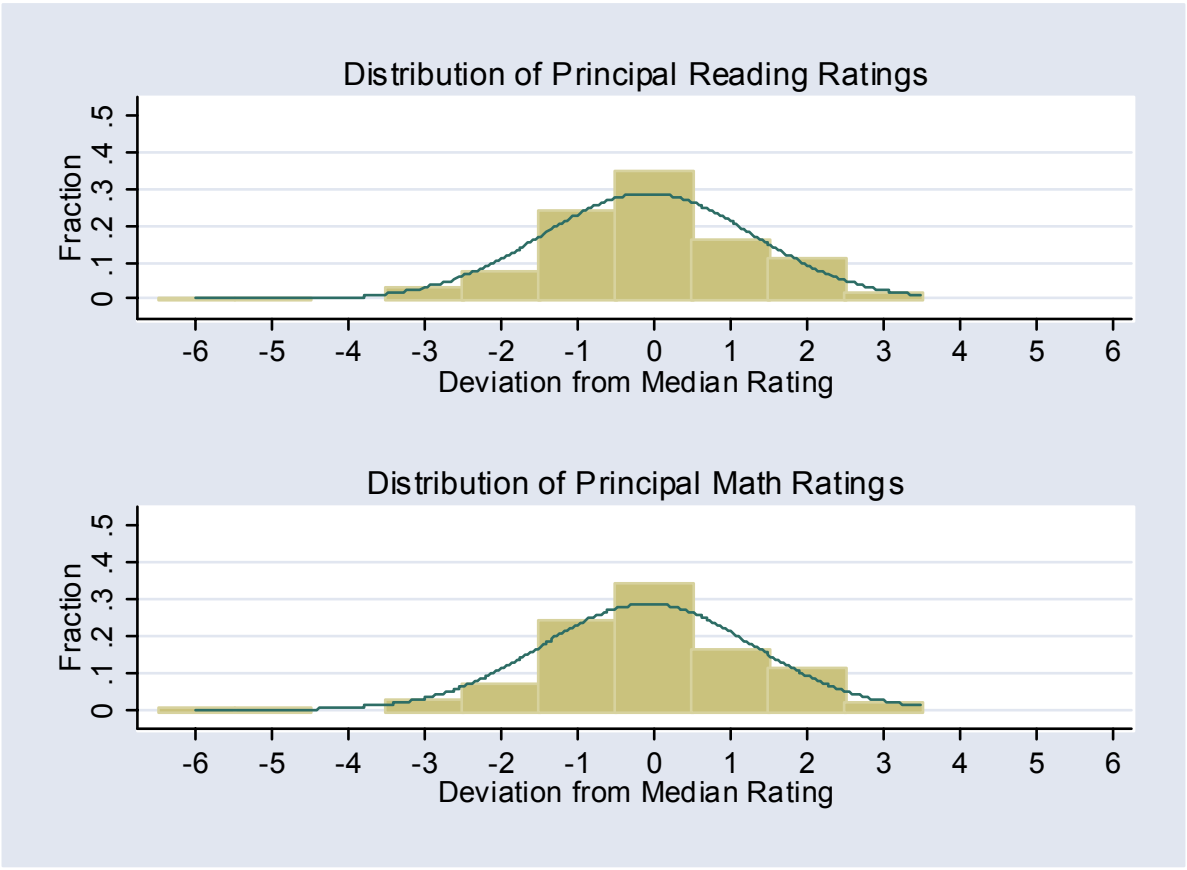
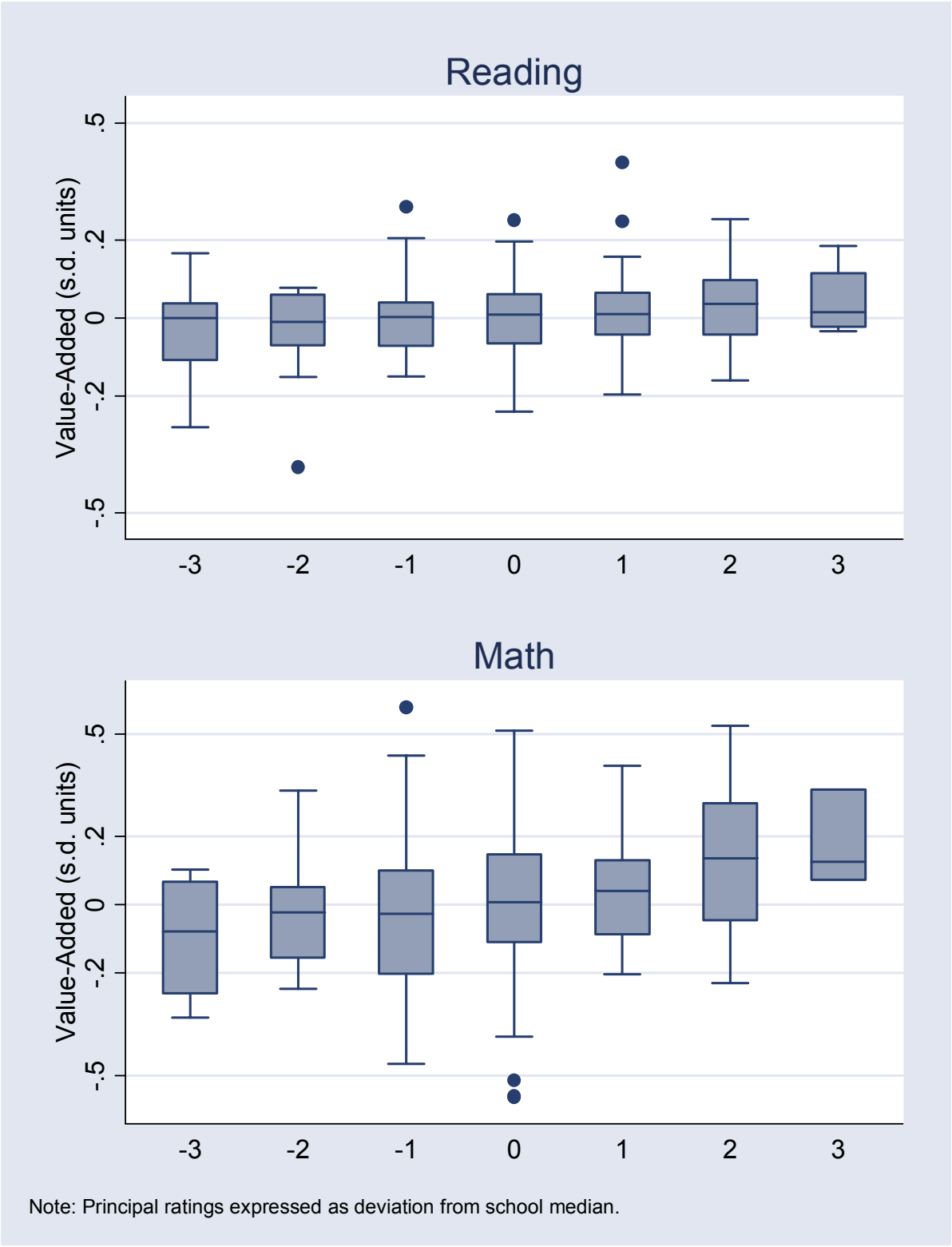
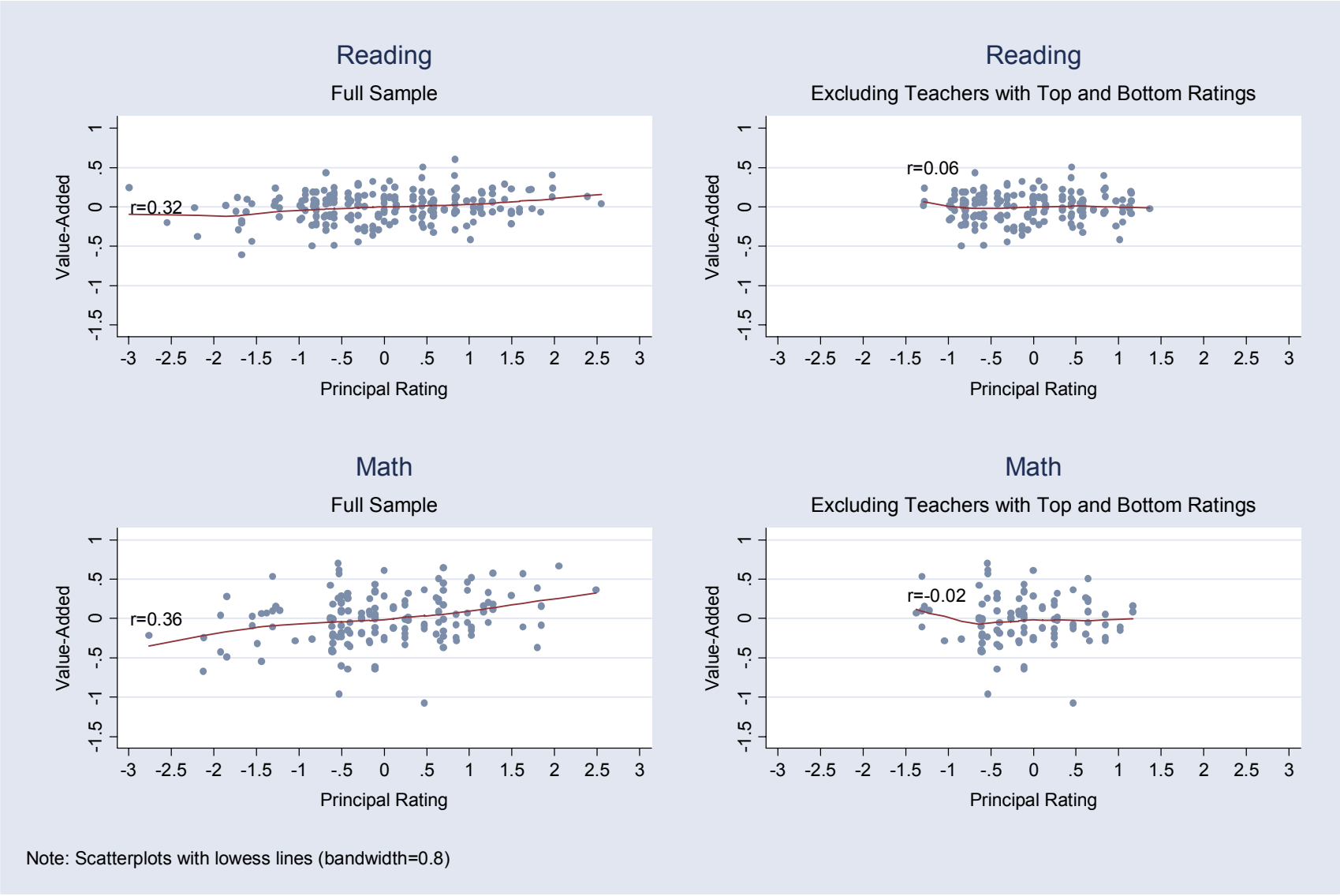


FIGURE II
THE DISTRIBUTION OF ESTIMATED TEACHER VALUE-ADDED BY PRINCIPAL RATING



Note: Principal ratings expressed as deviation from school median.

FIGURE III
ASSOCIATION BETWEEN ESTIMATED TEACHER VALUE-ADDED AND PRINCIPAL RATING



Note: Scatterplots with lowess lines (bandwidth=0.8)