

# Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education\*

Brian A. Jacob  
Harvard University and NBER

Lars Lefgren  
Brigham Young University

June 2007

---

\* 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](mailto: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](mailto:l-lefgren@byu.edu).

# Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education

## Abstract

In this paper, we examine how well principals can distinguish between more and less effective teachers. To put principal evaluations in context, we compare them with the traditional determinants of teacher compensation – education and experience – as well as value-added measures of teacher effectiveness based on student achievement gains. Moreover, we present “out-of-sample” predictions which mitigate concerns that the teacher quality and student achievement measures are determined simultaneously. We find that principals can generally identify teachers who produce the largest and smallest standardized achievement gains, but have far less ability to distinguish between teachers in the middle of this distribution.

JEL codes: I20, I28, J30, J70

"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."

Justice Potter Stewart (trying to define obscenity)

## I. Introduction

One of the most striking findings in recent education research involves the importance of teacher quality. A series of new papers have documented substantial variation in teacher effectiveness in a variety of settings, even among teachers in the same school (Aaronson, Barrow, and Sander 2007, Rockoff 2004, Hanushek et al. 2005). The differences in teacher quality are dramatic. For example, recent estimates suggest that the benefit of moving a student from an average teacher to one at the 85<sup>th</sup> percentile is comparable to a 33 percent reduction in class size (Rockoff 2004, Hanushek et al. 2005). The difference between having a series of very good teachers versus very bad teachers can be enormous (Sanders and Rivers 1996). At the same time, researchers have found little association between observable teacher characteristics and student outcomes – a notable exception being a large and negative first-year teacher effect (see Hanushek 1986 and 1997 for reviews of this literature and Rockoff 2004 for recent evidence on teacher experience effects).<sup>1</sup> This is particularly puzzling given the likely upward bias in such estimates (Figlio 1997, Rockoff 2004).

Private schools and most institutions of higher education implicitly recognize such differences in teacher quality by compensating teachers, at least in part, on the basis of ability (Ballou 2001, Ballou and Podgursky 2001). On the other hand, public school teachers have

---

<sup>1</sup> There is some limited evidence that cognitive ability, as measured by score on a certification exam for example, may be positively associated with teacher effectiveness, although other studies suggest that factors such as the quality of one's undergraduate institution are not systematically associated with effectiveness. For a review of the earlier literature relating student achievement to teacher characteristics, see Hanushek (1986, 1997). In recent work, Clotfelter et al. (2006) find teacher ability is correlated with student achievement while Harris and Sass (2006) find no such association.

resisted such merit-based pay due, in part, to a concern that administrators will not be able to recognize (and thus properly reward) quality (Murnane and Cohen 1986).<sup>2</sup> At first blush, this concern may seem completely unwarranted. Principals not only interact with teachers on a daily basis – reviewing lesson plans, observing classes, talking with parents and children – but also have ready access to student achievement scores. Prior research, however, suggests that this task might not be as simple as it seems. Indeed, the consistent finding that certified teachers are no more effective than their uncertified colleagues suggests that commonly held beliefs among educators may be mistaken.

In this paper, we examine how well principals can distinguish between more and less effective teachers, where effectiveness is measured by the ability to raise student math and reading achievement. In other words, do school administrators know good teaching when they see it? 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.

While there are several limitations to our analysis, which we describe in later sections of this article, our results suggest that policymakers should consider incorporating principal evaluations into teacher compensation and promotion systems. To the extent that principal judgments focus on identifying the best and worst teachers, for example to determine bonuses and teacher dismissal, the evidence presented here suggests that such evaluations would help promote student achievement. Principals can also evaluate teachers on the basis of a broader spectrum of educational outputs, including non-achievement outcomes valued by parents.

---

<sup>2</sup> Another concern, which we discuss below, involves favoritism or the simple capriciousness of ratings.

More generally, our findings inform the education production function literature, providing compelling evidence 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 makes a contribution to the empirical literature on subjective performance assessment by demonstrating the importance of accounting for estimation error in measured productivity and showing that the relationship between subjective evaluations and actual productivity can vary substantially across the productivity distribution.

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. Background**

### Prior Literature

The theoretical literature on subjective performance evaluation has focused largely on showing the conditions under which efficient contracts are possible (Bull 1987, MacLeod and Malcomson 1989). Prendergast (1993) and Prendergast and Topel (1996) show how the existence of subjective evaluation helps to explain several features of organizations such as the tendency of employers to agree with their employees. Recent work demonstrates that the compression and leniency in performance evaluations (relative to actual performance) often found in practice are features of the optimal contract between a risk-neutral principal and a risk-

averse agent when rewards are based on subjective performance evaluation (Levin 2003, MacLeod 2003).

The empirical literature on subjective performance measurement has focused largely on understanding the extent to which subjective supervisor ratings match objective measures of employee performance, and the extent to which subjective evaluations are biased. This research suggests that there is a relatively weak relationship between subjective ratings and objective performance (see Heneman 1986 and Bommer et al 1995 for good reviews), and that supervisor evaluations are indeed 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 (Alexander & Wilkins, 1982; Bolino & Turnley, 2003; Heneman, Greenberger & Anonyuo, 1989; Lefkowitz, 2000; Wayne & Ferris, 1990; Varma & Stroh, 2001).<sup>3</sup>

The studies most closely related to the present analysis examine the subjective evaluations given to elementary school teachers by their principals. A collection of studies in the education literature report quite small correlations between principal evaluations and student achievement, although these studies are generally based on small, non-representative samples, do not account properly for measurement error and rely on objective measures of teacher performance that are likely biased (Medley and Coker 1987 and Peterson 1987, 2000).<sup>4</sup> The best study on this topic examines the relationship between teacher evaluations and student achievement among second and third graders in the New Haven public schools (Murnane 1975). Conditional on prior student test scores and demographic controls, the author found that teacher

---

<sup>3</sup> Prendergast (1999) observes that such biases create an incentive for employees to engage in inefficient “influence” activities. Wayne and Ferris (1990) provide some empirical support for this hypothesis.

<sup>4</sup> 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).

evaluations were significant predictors of student achievement, although the magnitude of the relationship was only modest.<sup>5</sup>

### Conceptual Framework

In order to provide a basis for interpreting the empirical findings in this paper, it is useful to consider how principals might form opinions of teachers. Given the complexity of principal belief formation, and the limited objectives of this paper, we do not develop a formal model. Rather, we describe the sources of information available to principals and how they might interpret the signals they receive.

Each year principals receive a series of noisy signals of a teacher's performance, stemming from three main sources: (1) formal and informal observations of the teacher working with students and/or interacting with colleagues around issues of pedagogy or curriculum; (2) reports from parents, either informal assessments or formal requests to have a child placed with the teacher (or not placed with the teacher); and (3) student achievement scores. Principals will differ in their ability and/or inclination to gather and incorporate information from these sources, and in the weight that they place on each of the sources. A principal will likely have little information on first-year teachers, particularly at the point when we surveyed principals; namely, in February, before student testing took place and before parents began requesting specific teachers for the following year.

---

<sup>5</sup> 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 one 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) present similar results for a subset of high poverty schools in Los Angeles. They 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).

Principals may differ with respect to the level of sophistication with which they collect information and interpret the signals they receive. For example, principals may be aware of the level of test scores in the teacher's classroom but be unable to account for differences in classroom composition. In this case, principal ratings might be more highly correlated to the level of test scores than to teacher value-added if little information besides test scores was used to construct ratings. Also, principals might vary in how they deal with the noise inherent in the signals they observe. A naïve principal might simply report the signal she observes regardless of the variance of the noise component. A more sophisticated principal, on the other hand, might act as a Bayesian, down-weighting noisy signals.<sup>6</sup> Finally, it seems likely that a principal's investment in gathering information on and updating beliefs about a particular teacher will be endogenously determined by a variety of factors, including the initial signal the principal observes as well as the principal's assessment regarding how much a teacher can benefit from advice and training.

Ultimately, in this paper we limit our examination to the accuracy of principal ratings regardless of the strategies principals use to construct them. In an early version of this paper, we presented evidence inconsistent with principals acting as perfect Bayesians. Based on this work, we concluded that the actual algorithms used by principals to form their opinions of teacher effectiveness are likely to be quite complex and highly variable across individual administrators. For this reason, we defer to future research the construction and testing of behavioral models of principal evaluations.

---

<sup>6</sup> Indeed, a simple model of principals as perfect Bayesians would generate strong implications regarding how the accuracy of ratings evolves with the time a principal and teacher spend together. For example, this type of model would imply that the variance of principal ratings *increases* over time (assuming that the variance of true teacher ability does not change over time).

### III. Data

The data for this study comes from a mid-size school district located in the western United States. 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 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 (49<sup>th</sup> 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.<sup>7</sup> We exclude kindergarten and first grade teachers because achievement exams are not available for these students.<sup>8</sup>

Our sample consists of 201 teachers in grades two through six. Like the students, the teachers in our sample are fairly representative of elementary school teachers nationwide. Only

---

<sup>7</sup> We exclude non-core teachers such as music teachers, gym teachers and librarians. Note, however, that in calculating teacher value-added measures, we use all student test scores from 1997-98 through 2004-05.

<sup>8</sup> 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.

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 6 percent attended a school out of state. 16 percent of teachers have a masters degree or higher, and the vast majority of teachers are licensed in either early childhood education or elementary education. Finally, 7 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. 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 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*.

### 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.<sup>9</sup> 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

---

<sup>9</sup> 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.

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 within the school for that same item. It appears that principal ratings within a given school are roughly normally distributed with five to six relevant categories. Perhaps not surprisingly, the principal responses to some individual survey items are highly correlated (e.g., the correlation between teacher organization and classroom management exceeds 0.7), while others are less highly correlated (e.g., the correlation between role model and relationship with colleagues is less than 0.4). 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.

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

The primary challenge in estimating measures of teacher effectiveness using student achievement data involves the potential for non-random assignment of students to classes. 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, Barrow and Sander 2007, 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.

For our baseline specification, we use a panel of student achievement data from 1997-98 through 2002-03 to estimate the following model:

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

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.  $\psi_t$  and  $\phi_k$  are a set of year and school fixed effects, respectively. Teacher  $j$ 's contribution to value-added is captured by the  $\delta_j$ 's.  $\alpha_{jt}$  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).  $\varepsilon_{ijkt}$  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 method suggested by Moulton (1990).<sup>10</sup>

---

<sup>10</sup> Another possibility would be to use cluster-corrected standard errors. However, the estimated standard errors behave very poorly for teachers who are in the sample for a small number of years. It is also possible to estimate a

It is worth noting that our value-added model differs from standard practice in one respect. Given that ratings are normalized to have equal mean and variance for each principal, a value-added indicator that measures effectiveness relative to a district rather than school average will be biased downward.<sup>11</sup> 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.<sup>12</sup>

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. To account for unobservable, time-invariant student characteristics (e.g., motivation or parental involvement), we estimate value-added models that utilize achievement normalized *gains* as the outcome and include student fixed effects (but not prior achievement measures) as controls.<sup>13</sup> As we show

---

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. The intra-class correlation coefficients calculated as part of the Moulton procedure are roughly .06 in reading and .09 in mathematics.

<sup>11</sup> Typical value added models simply contain school fixed effects that identify teacher quality relative to all teachers (or some omitted teacher) in the district. The comparison of teachers across schools is identified by both covariates in the model as well as the fact that the same teachers switch schools during the sample period. To see that the use of district-relative value-added measures will lead to a downward bias, consider the possibility that teachers in certain schools have systematically higher value-added measures than teachers in other schools. In the extreme, value-added could be perfectly sorted across schools so that all of the teachers in the “best” school have higher value-added than teachers in the “second best” school, and teachers in this second best school all have higher value-added scores than the teachers in the third best school, etc. However, because we have normalized principal ratings within school, this means that half of the teachers in the best school will, by construction, have “below average” subjective ratings. The fact that the subjective and objective ratings are measured on different scales in this example essentially introduces measurement error that will attenuate our correlations.

<sup>12</sup> 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).

<sup>13</sup> Following Hanushek et al. (2005) and Reback (2005), we normalized achievement gains by student prior ability. Specifically, we divide students into  $q$  different quantiles based on their prior achievement score,  $y_{ijkt-1}$ , and then

below, our results are robust to models that include student fixed effects. We do not include such models as our baseline; however, as fewer students contribute to the identification of teacher value-added with student fixed effects (e.g. sixth grade students in the first year of our data and second grade students in the last are omitted). Also, the correlation of the estimation error across individual teacher value-added measures is higher than in our baseline specifications complicating estimation. For more details on the value-added models used in this paper, see Jacob and Lefgren (2005a).

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 0.12 in reading and 0.26 in math, which appears roughly consistent with recent literature on teacher effects.<sup>14</sup> Because the dependent variable is a state-specific, criterion-referenced test that we have normalized within grade-year for the district, 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. To provide a better sense of the magnitude of these effects, 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

---

calculate the mean and standard deviation of achievement *gains* ( $g_{ijkt} = y_{ijkt} - y_{ijkt-1}$ ) for each quantile separately, which we denote as  $\mu_g^q$  and  $\sigma_g^q$  respectively. We then create standardized gain scores that are mean zero and unit standard deviation *within* each quantile:  $G_{ijkt}^q = (g_{ijkt}^q - \mu_g^q) / \sigma_g^q$ . This ensures that each student's achievement gain is compared to the improvement of comparable peers.

<sup>14</sup> 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. Examining high school students, Aaronson, Barrow and Sander (2007) find that a one standard deviation improvement in teacher quality leads to a .20 improvement in math performance over the course of a year.

mean would result in roughly a 2-3 percentile point increase in test scores (an increase of roughly 0.07-0.10 standard deviation units).

#### IV. Empirical Strategy

The primary objective of our analysis is to determine how well principals can distinguish between more and less effective teachers as measured by student achievement gains on standardized exams. This section outlines the strategies we employ to address this issue. Our empirical methods take into account estimation error in our value-added measures and the ordinal nature of principal ratings.

Perhaps the most straightforward measure of association between the subjective principal assessments and the objective value-added measures is a simple correlation. However, the estimation error in our value-added measures – 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) – will lead us to understate the correlation between the principal ratings and the value-added indicators.<sup>15</sup> To see this, note that the observed value-added can be written as  $\hat{\delta}^{OLS} = \delta + e$  where  $\delta$  is the true fixed effect and  $e$  represents 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:

---

<sup>15</sup> 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.

$$(2) \quad \text{Corr}(\hat{\delta}^p, \hat{\delta}^{OLS}) = \frac{\text{Cov}(\hat{\delta}^p, \hat{\delta}^{OLS})}{\sqrt{\text{Var}(\hat{\delta}^p)\text{Var}(\hat{\delta}^{OLS})}} = \frac{\text{Cov}(\hat{\delta}^p, \delta)}{\sqrt{\text{Var}(\hat{\delta}^p)[\text{Var}(\delta) + \text{Var}(e)]}} < \text{Corr}(\hat{\delta}^p, \delta)$$

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:  $\text{Var}(\delta) = \text{Var}(\hat{\delta}^{OLS}) - \text{Var}(e)$ .<sup>16</sup>

Then we can then simply multiply the observed correlation,  $\text{Corr}(\hat{\delta}^p, \hat{\delta}^{OLS})$ , by  $\frac{\sqrt{\text{Var}(\hat{\delta}^{OLS})}}{\sqrt{\text{Var}(\delta)}}$  to obtain the adjusted correlation. We obtain the standard errors using a bootstrap.<sup>17</sup>

Note that this adjustment assumes that a principal's rating is unrelated to the error of our OLS estimate of teacher effectiveness. Specifically, we assume that the numerator in equation (2) can be rewritten as follows:  $\text{Cov}(\hat{\delta}^p, \hat{\delta}^{OLS}) = \text{Cov}(\hat{\delta}^p, \delta) + \text{Cov}(\hat{\delta}^p, e)$ . 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

---

<sup>16</sup> 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. Estimation 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.

<sup>17</sup> For our baseline specifications we perform 1000 iterations. For the robustness and heterogeneity checks we perform 500 iterations. We also perform a principal-level block bootstrap. The inference from this procedure (shown below) is similar to our baseline results.

model does – that is, conditioning on a variety of covariates), the correlation we calculate will be biased upwards.<sup>18</sup>

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.<sup>19</sup> To 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).<sup>20</sup> For more information on our calculation of the EB estimates, see Jacob and Lefgren (2005a).

While the correlation between objective and subjective performance measures is a useful starting point, it has several limitations. Most importantly, the principal ratings may not have a cardinal interpretation, which would make the correlation impossible to interpret. For example, the difference between a 6 and 7 rating may be greater or less in absolute terms than the difference between a 9 and 10. Second, correlations are quite sensitive to outliers. Third, while we have normalized the principal ratings to ensure that each principal's ratings have the same variance, it is possible that the variance of value-added differs across schools. In this case, stacking the data could provide misleading estimates of the average correlation between principal

---

<sup>18</sup> 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. Similarly, the correlations may be biased downward if principals assign teachers to classrooms in a compensatory way – that is, principals assign more effective teachers to classrooms with more difficult students. In this case, our value-added measures will be attenuated (biased toward zero), which will reduce the correlation between our subjective and objective measures.

<sup>19</sup> If the value-added measure is used as a dependent variable, it will lead to less precise 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.

<sup>20</sup> 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).

ratings and value-added within the district. 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 estimate a non-parametric measure of the association between ratings and productivity. Specifically, we calculate the probability that principal can correctly identify teachers in the top or 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.<sup>21</sup> 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 below. Appendix A provides a more detailed description of this simulation.

A final concern with both the parametric and non-parametric measures of association described above is that our objective measure of teacher effectiveness is constructed from student test scores which have no natural units. Exam scores in this district are reported in terms of the percent of items answered correctly. We do not have access to individual test items, which makes it impossible to develop performance measures that account for the difficulty of the test items, as is commonly done in Item Response Theory (IRT) analyses.<sup>22</sup> Furthermore, until very recently the test was not standardized against a national population. However, to test the

---

<sup>21</sup> The normality of teacher ratings around the point estimate is an implication of the Central Limit Theorem.

<sup>22</sup> The Core Exam is the exam that is administered to all grades and broadly reported. It is therefore likely that principals would focus on this measure of achievement as opposed to some other exam, such as the Stanford Achievement Test, which is only administered in the third (with limited participation), fifth, and eighth grades.

robustness of our results, we develop value-added measures that use transformations of the percent correct score, including a student's percentile rank within his year and grade, the square of the percent correct and the natural logarithm of the percent correct. Moreover, the district categorizes students into four different proficiency categories on the basis of these exam scores, and in some specifications we use these proficiency indicators as outcomes for the creation of value-added measures. As we show below, our results are robust to using these alternative value-added measures.

## **V. Can Principals Identify Effective Teachers?**

Having outlined our strategy for estimating the relationship between principal evaluations and teacher effectiveness, we will now present our empirical findings. We will also compare the usefulness of principal assessments in *predicting* future teacher effectiveness relative 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?

Table III 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 0.18 and 0.25 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 0.29 and 0.32 respectively. It is important to emphasize that these correlations are not based 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 measures are identifying the same underlying construct, they should not be biased downward as in the case with many prior studies of subjective performance evaluation.<sup>23</sup> The positive and significant correlations indicate that principals do have some ability to identify this dimension of teacher effectiveness. As shown below, these basic results are robust to a wide variety of alternative specifications and sensitivity analyses.

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 than the correlation with between principal ratings and teacher value-added (0.55 versus 0.29). This suggests that principals may base their ratings at least partially on a naïve recollection of student performance in teacher's class—failing to account for differences in classroom composition. Another reason may be that principals focus on their most recent observations of teachers. In results not shown here, we find 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.<sup>24</sup>

---

<sup>23</sup> Bommer 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 is no higher than with an overall principal rating.

<sup>24</sup> This exercise suggests that in assigning teacher ratings, principals might focus on their most recent observations of a teacher instead of incorporating information on all past observations of the individual. 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

To explore principals' ability to identify the very best and worst teachers, Table IV shows the estimates of the percent of teachers that a principal can correctly identify in the top (bottom) group within his or her school. 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 64 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).<sup>25</sup>

The second and third panels in Table IV 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 51 percent of teachers as being better than the median (including those in the top category), 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.<sup>26</sup>

---

value-added measures that incorporate a time-varying teacher experience measure. As shown in our sensitivity analysis, we obtain comparable results when we use this measure.

<sup>25</sup> We used the delta method to compute the relevant standard errors. A block bootstrap at the principal level yielded even smaller p-values than those reported in the tables.

<sup>26</sup> An alternative explanation for this finding is that principal ratings are noisy, so the difference in value-added between two teachers with principal evaluations that differ by one point is much smaller than the difference in value-added between teachers whose ratings differ by two or more points. If this were true, we might expect principals to appear capable of identifying top teachers, not because they can identify performance in tails of the distribution, but rather because top teachers have much higher principal evaluations than the remainder of the

Figure II 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 0.02 and -0.04 in reading and math respectively (neither of which is significantly different than zero).

### Robustness Checks

In this section, we outline potential concerns regarding the validity of our estimates and attempt to provide evidence regarding the robustness of our findings. One reason that principals might have difficulty distinguishing between teachers in the middle is that the distribution of teacher value-added is highly compressed. This is not the case, however. Even adjusting for estimation error, the standard deviation of value-added for teachers outside of the top and bottom categories is .10 in reading and .25 in math compared with standard deviations of .12 and .26 respectively for the overall sample.

Another possible concern is that the lumpiness of the principal ratings might reduce 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

---

sample. A comparison of teachers not in the top or bottom groups but who had the same difference in absolute ratings might yield similar results. To investigate this possibility, we examined whether the difference in principal ratings predicted which teacher had higher estimated value-added. We found that, controlling for whether teachers had received the top or bottom rating, the difference in principal ratings between two teachers was not predictive of which teacher had higher estimated value-added. This suggests that our findings do indeed reflect a principal's ability to identify teachers in the tails of performance distribution.

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.<sup>27</sup> In addition, as shown in row 5, we obtain comparable correlations when we exclude principals with extremely lumpy ratings.

Table V presents a variety of additional sensitivity analyses. For the sake of brevity, we only show the correlations between principal ratings and value-added, adjusting for the estimation error in value-added. In row 1, we present the baseline estimates. In the first panel (rows 2 through 5), we examine the sensitivity of our findings to alternative samples. We find similar results when we examine only grades for which we can examine math as well as reading. The same is true if we exclude schools with extremely lumpy principal ratings, or first year teachers. As discussed earlier, the correlation between ratings and teacher value-added is much lower for teachers with ratings outside of the top and bottom.

In the next panel (rows 6 to 12), we investigate the concern that our findings may not be robust to the measure of student achievement used to calculate value-added. As explained earlier, while principals were likely to focus on the Core exam, it is possible that they weigh the

---

<sup>27</sup> 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.

improvement of high and low ability students in a way that does not correspond to the commonly reported test metric. In rows 6-8, we examine measures of achievement that focus on students at the bottom of the test distribution. These measures include the log of the percent correct, achieving above the proficiency cutoff, and the achievement of students with initial achievement below the classroom mean. We find results comparable to our baseline. In rows 9-10, we show results in which we place greater weight on high achieving students. These include percent correct squared and the achievement of students with initial achievement above the classroom mean, and find somewhat smaller correlations. Overall, we believe these results are consistent with our baseline findings. However, there does seem to be some suggestive evidence that principals may place more weight on teachers bringing up the lowest-achieving students. For example, the correlations using the teacher value-added scores based on the achievement of students in the bottom half of the initial ability distribution are larger than those based on the students in the top half of the ability distribution (though the differences are not significant at conventional levels). On the other hand, the correlations that use value-added measures based on proficiency cutoffs (which should disproportionately reflect impacts on students with lower initial ability) are smaller than our baseline results and not statistically different than zero. Of course, this might also be due to the fact that this binary outcome measure inherently contains less information than a continuous achievement score.

In row 11, we examine results in which the achievement measure used is the students percentile in the grade-year distribution of test scores within the district. Finally, in row 12, we examine gain scores that are normalized so that student gains have unit standard deviation within each decile of the initial achievement distribution. Both specifications yield results comparable to the baseline.

In the third panel (rows 13 to 18), we examine the robustness of our estimates to alternative estimation strategies for the value-added, and find results comparable to our baseline.<sup>28</sup> Row 19 demonstrates that our inferences are robust to calculating standard errors using a block bootstrap that clusters observations by school. Row 20 addresses the potential concern that the variance of teacher value-added may not be constant across principals, in which case stacking the data may be inappropriate. Here we estimate the unadjusted correlations separately for each principal and take the simple average of them. We find that averaging the unadjusted correlations across the principals yields an estimate similar to our baseline unadjusted correlation.

#### Heterogeneity of Effects

It is possible that principals are able to gauge the performance of some teachers more effectively than others. Alternatively, some principals may be more effective than others in evaluating teacher effectiveness. In Table VI, we examine this possibility. The first row shows the baseline estimates.

In the first panel, we examine how the correlation between principal rating and teacher value-added varies with teacher characteristics. It does not appear that the correlation between ratings and teacher value-added varies systematically with teacher experience, the duration the principal has known the teacher, or grade taught. The standard errors are generally too large, however, to draw strong conclusions.

In the second panel, we consider whether some principals are more capable of identifying effective teachers than others. Principals who have been at their schools less than four years, are

---

<sup>28</sup> One exception is that the correlations that use only 2002-03 achievement data are smaller than the baseline correlations, and are not statistically different than zero (although they are not statistically different than the baseline correlations either).

male, and identify themselves as confident in their ability to assess teacher effectiveness appear to rate teaching ability more accurately. The observed differences, however, are never significant at the 5 percent level.

### A Comparison of Alternative Teacher Assessment Measures

To put the usefulness of principal ratings into perspective, it is helpful to compare them to alternative metrics of teacher quality. These include 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.<sup>29</sup> This exercise serves an additional purpose as well. To the extent that principals observe past test scores, their ratings may be correlated to the estimation error of the value-added estimates. In this section, we overcome this concern by examining student achievement *subsequent* to the principal evaluations.

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 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

---

<sup>29</sup> 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.

the regressions, we use empirical Bayes estimates of the value-added (see Morris, 1983). 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, limiting our sample to 162 teacher observations for reading and 118 teacher observations for math.<sup>30</sup> 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.<sup>31</sup>

Table VII 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. In contrast, teachers with advanced degrees have students that score roughly 0.11 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

---

<sup>30</sup> 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.

<sup>31</sup> 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.

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 0.07 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 half again as that on the overall principal ratings.<sup>32</sup> 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 the 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 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 predictor of student achievement. The results for math, shown in column 12, are even stronger.

---

<sup>32</sup> 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.

## VI. Conclusions

In this paper, we examine principals' ability to identify teachers' ability to increase reading and math achievement. We build on prior literature by using principal ratings that are aligned with the objective metric under examination. We also take into account measurement error in the objective metric and the fact that principal ratings are categorical and may not have a cardinal interpretation. We find that principals are generally effective at identifying the very best and worst teachers. On average, however, they are not able to distinguish teachers in the middle of the achievement distribution. There is suggestive evidence that principals may be more concerned with or aware of the achievement of low-ability students than of high ability students, and may rely on achievement levels rather than value-added to assess teachers. Principal ratings are also a significant predictor of future student achievement, though they perform worse than empirical measures of teacher effectiveness.

Our results suggest that policymakers should consider incorporating principal evaluations into teacher compensation and promotion systems. Comparing principal assessments to the traditional determinants of teacher compensation – education and experience – we find that subjective principal assessments of teachers are a substantially better predictor of future student achievement. 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 at identifying the best and worst teachers. To the extent that principal judgments focus on identifying the best and worst teachers, for example to determine bonuses and teacher dismissal, the evidence presented here suggests that such evaluations would help promote student achievement. Moreover, principal assessment has the potential to mitigate some of the concerns regarding strategic behavior on the part of teachers to improve test scores without increasing

actual knowledge (see Jacob and Levitt, 2003).<sup>33</sup> 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. Also, in other work we show that principal ratings are correlated to other educational outputs valued by parents, such as student satisfaction (see Jacob and Lefgren, forthcoming).

While principal evaluations seem promising, there are several important reasons to be cautious in using these results to shape teacher compensation and promotion systems. First, the inability of principals to distinguish between a broad middle-range of teacher quality suggests that one should not rely on principals for fine grained performance determinations as might be required under certain merit pay policies. Second, our analysis takes place in a context where principals were not explicitly evaluated on the basis of their ability to identify effective teachers.<sup>34</sup> 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. Favoritism toward particular teachers by school administrators long has been a concern among teachers, and Jacob and Lefgren (2005a) find some tentative evidence that principals may indeed engage in such behavior. Third, 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

---

<sup>33</sup> 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).

<sup>34</sup> 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.

substantial impact on student outcomes unless it were accompanied by meaningful consequences.<sup>35</sup>

More broadly, our findings provide compelling evidence 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. This should provide some hope to those attempting to characterize the behaviors associated with effective teaching, and ultimately improve education for all students.

---

<sup>35</sup> For examples of studies that examine accountability programs within education, see Jacob 2005, Kane and Staiger (2002), Figlio and Rouse (2006), Figlio and Winicki (2005).

## References

- Aaronson, Daniel, Lisa Barrow and William Sander (2007). "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25(1), 95-135.
- 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.
- Ballou, Dale (2001). "Pay for performance in public and private schools." *Economics of Education Review* 20: 51-61.
- Ballou, Dale and Michael Podgursky (2001). "Teacher Compensation: The Case for Market-Based Pay." *Education Matters* 1(1), 16-25.
- 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.
- Bull, Clive (1987). "The Existence of Self-Enforcing Implicit Contracts." *Quarterly Journal of Economics* 102(1): 147-160.
- Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor, "Teacher-Student Matching and the Assessment of Teacher Effectiveness," NBER Working Paper No. 11936 (2006).
- Figlio, David N. (1997). "Teacher Salaries and Teacher Quality." *Economics Letters*. August.
- Figlio, David N and Maurice E. Lucas (2004). "Do High Grading Standards Affect Student Performance?" *Journal of Public Economics* 88(9-10): 1815-1834.
- Figlio, David N. and Cecilia Elena Rouse (2006). "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics* 90(1-2): 239-255.

- Figlio, David N. and Joshua Winicki (2005). "Food For Thought: the Effects of School Accountability on School Nutrition." *Journal of Public Economics* 89(2-3): 381-394.
- 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. (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 No. 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.
- Harris, Douglas and Timothy Sass (2006). "The Effects of Teacher Training on Teacher Value-Added" Working Paper, Florida State University.
- 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 (2005a). "Principals as Agents: Subjective Performance Measurement in Education." National Bureau of Economic Research Working Paper No. 11463.
- Jacob, Brian A. and Lars Lefgren (2005b). "What Do Parents Value in Education? An Empirical Investigation of Parents' Revealed Preferences for Teachers." National Bureau of Economic Research Working Paper No. 11494.
- 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. and D. Staiger (2002). "Volatility in School Test Scores: Implications for Test-Based Accountability Systems." In: Ravitch, D. (Ed.), *Brookings Papers on Education Policy 2002*. Washington, DC: Brookings Institution Press.

- Lefkowitz, Joel (2000). "The Role of Interpersonal Affective Regard in Supervisory Performance Ratings: A Literature Review and Proposed Causal Model." *Journal of Occupational and Organizational Psychology* 73: 67-85.
- Levin, Jonathan (2003). "Relational Incentive Contracts." *American Economic Review* 93(3): 835-857.
- MacLeod, W. Bentley (2003). "Optimal Contracting with Subjective Evaluation." *American Economic Review* 93(1): 216-240.
- MacLeod, W. Bentley and J. Malcomson (1989). "Implicit Contracts, Incentive Compatibility and Involuntary Unemployment." *Econometrica* 56(2): 447-480.
- 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. (1990). "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *Review of Economics and Statistics* 72(2): 334-338.
- Murnane, Richard (1975). *The Impact of School Resources on the Learning of Inner-City Children*. Cambridge, MA: Ballinger Publishing Company.
- Murnane, R. J. and D. K. Cohen (1986). "Merit pay and the evaluation problem: why most merit pay plans fail and few survive." *Harvard Educational Review* 56 (1): 1-17.
- Prendergast, Canice (1993). "The Role of Promotion in Inducing Specific Human Capital Acquisition." *Quarterly Journal of Economics* 108(2): 523-534.
- Prendergast, Candice (1999). "The Provision of Incentives in Firms." *Journal of Economic Literature* 37(1): 7-63.
- Prendergast, Canice and Robert Topel (1996). "Favoritism in Organizations." *Journal of Political Economy* 104(5): 958-975.
- Peterson, Kenneth D. (2000). *Teacher Evaluation: A Comprehensive Guide to New Directions and Practices* (2nd 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.
- Sanders, W. L. and Rivers, J. C. (1996). "Cumulative and residual effects of teachers on future student academic achievement." University of Tennessee Value-Added Research and Assessment Center.
- 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.
- 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:

### 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  $\delta_j$ . Our goal is to calculate the following probabilities:

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

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

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)

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

(D4)

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

(D5)

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

(D6)

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

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

$$\begin{aligned} \Pr(\hat{\delta}_j = t | \delta_j = t) &= \Pr(\hat{\delta}_j = t | \delta_j = t, \hat{\delta}_j^P = t) \\ \Pr(\hat{\delta}_j = t | \delta_j = b) &= \Pr(\hat{\delta}_j = t | \delta_j = b, \hat{\delta}_j^P = t) \end{aligned}$$

While we do not believe this is strictly true, it should not substantially bias our estimates.

We also know the following identities are true:

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

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

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

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

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

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

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

$$\begin{aligned}
\text{(D12)} \quad \Pr(\delta_j = t \mid \hat{\delta}_j^P = t) &= \frac{\Pr(\hat{\delta}_j = t \mid \hat{\delta}_j^P = t) - \Pr(\delta_j = b \mid \hat{\delta}_j = t) \frac{\Pr(\hat{\delta}_j = t)}{\Pr(\delta_j = b)}}{\Pr(\delta_j = t \mid \hat{\delta}_j = t) - \Pr(\delta_j = b \mid \hat{\delta}_j = t) \frac{\Pr(\hat{\delta}_j = t)}{\Pr(\delta_j = b)}}
\end{aligned}$$

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:

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

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

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

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

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  $\text{Var}(\hat{\delta}_j)$ . 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  $\Pr(\hat{\delta}_j = t) = \Pr(\delta_j = t)$  and  $\Pr(\hat{\delta}_j = b) = \Pr(\delta_j = b)$  directly from our original data. In many cases, for example, because we are interested in the top versus bottom quantiles, we know that  $\Pr(\hat{\delta}_j = t) = \Pr(\delta_j = t) = \Pr(\hat{\delta}_j = b) = \Pr(\delta_j = b)$ , 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):

$$\begin{aligned}
 \Pr(\delta_j = b \mid \hat{\delta}_j^p = b) &= \frac{\Pr(\hat{\delta}_j = b \mid \hat{\delta}_j^p = b) - \Pr(\hat{\delta}_j = b \mid \delta_j = t)}{\Pr(\hat{\delta}_j = b \mid \delta_j = b) - \Pr(\hat{\delta}_j = b \mid \delta_j = t)} \\
 \text{(D17)} \quad &= \frac{\Pr(\hat{\delta}_j = b \mid \hat{\delta}_j^p = b) - \Pr(\delta_j = t \mid \hat{\delta}_j = b) \frac{\Pr(\hat{\delta}_j = b)}{\Pr(\delta_j = t)}}{\Pr(\delta_j = b \mid \hat{\delta}_j = b) - \Pr(\delta_j = t \mid \hat{\delta}_j = b) \frac{\Pr(\hat{\delta}_j = b)}{\Pr(\delta_j = t)}}
 \end{aligned}$$

TABLE I  
SUMMARY STATISTICS

<i>Student Characteristics</i>	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
<i>Teacher Characteristics</i>	Mean (s.d.)
Male	0.16 (0.36)
Age	41.9 (12.5)
Untenured	0.17 (0.38)
Experience	11.9 (8.9)
Fraction with 10-15 Years Experience	0.19 (0.40)
Fraction with 16-20 Years Experience	0.14 (0.35)
Fraction with 21+ Years Experience	0.16 (0.37)
Years working with principal	4.8 (3.6)
BA Degree at in state (but not local) college	0.10 (0.30)
BA Degree at out of state college	0.06 (0.06)
MA Degree	0.16 (0.16)
Any additional endorsements	0.20 (0.40)
Any additional endorsements in areas other than ESL	0.10 (0.31)
Licensed in more than one area	0.26 (0.44)
Licensed in area other than ECE or EE	0.07 (0.26)
2 <sup>nd</sup> Grade	0.23 (0.42)
3 <sup>rd</sup> Grade	0.21 (0.41)
4 <sup>th</sup> Grade	0.20

	(0.40)
5 <sup>th</sup> Grade	0.18
	(0.38)
6 <sup>th</sup> Grade	0.18
	(0.39)
Mixed grade classroom	0.07
	(0.26)
Two teachers in the classroom	0.05
	(0.22)
<hr/>	
Number of teachers	201
Number of principals	13
<hr/>	

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 <sup>th</sup> Percentile	90 <sup>th</sup> 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  
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		Diff: Reading – Math (Math Sample Only) (5)
	Unadjusted (1)	Adjusted (2)	Unadjusted (3)	Adjusted (4)	
(1) Using baseline specification for creating value-added measure	0.18* (0.07)	0.29* (0.10)	0.25* (0.08)	0.32* (0.10)	0.00 (0.11)
(2) Using average student achievement (levels) as the value-added measure	0.35* (0.05)	0.55* (0.09)	0.28* (0.08)	0.37* (0.12)	0.19 (0.13)
(3) Difference: (1) – (2)		-0.26* (0.08)		0.05 (0.07)	

Notes: Number of observations for reading and math is 201 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 IV  
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 <b>top rating</b> from the principal was the top teacher according to the value-added measure (standard error)	0.52 (0.17)	0.64 (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	2.30 (0.02)	3.06 (0.00)
Conditional probability that a teacher who received a rating <b>above the median</b> from the principal was above the median according to the value-added measure (standard error)	0.51 (0.10)	0.57 (0.14)
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	1.79 (0.07)	2.41 (0.02)
Conditional probability that a teacher who received a rating <b>below the median</b> from the principal was below the median according to the value-added measure (standard error)	0.48 (0.09)	0.44 (0.12)
Null hypothesis (probability expected if principals randomly assigned teacher ratings)	0.35	0.26
Z-score (p-value) of test of difference between observed and null	1.40 (0.16)	1.59 (0.11)
Conditional probability that the teacher(s) who received the bottom rating from the principal was the <b>bottom teacher(s)</b> according to the value-added measure (standard error)	0.44 (0.21)	0.60 (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	1.70 (0.09)	2.76 (0.01)

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

TABLE V  
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	0.18* (0.07)	0.29* (0.10)	0.25* (0.08)	0.32* (0.10)
	<i>Alternative Samples</i>				
2	For math sample	0.23* (0.08)	0.32* (0.11)	--	--
3	Exclude teachers who were in their first year in 2002-03	0.19* (0.07)	0.26* (0.10)	0.28* (0.08)	0.35* (0.10)
4	Exclude teachers with top or bottom principal rating	0.02 (0.07)	0.03 (0.13)	-0.04 (0.10)	-0.05 (0.14)
5	Exclude schools with extreme lumpiness in principal ratings	0.15* (0.07)	0.26* (0.13)	0.24* (0.09)	0.32* (0.12)
	<i>Alternative Measures of Student Achievement</i>				
6	Test outcome is natural log of percent correct—greater weight on low achieving students	0.19* (0.07)	0.33* (0.13)	0.22* (0.08)	0.28* (0.10)
7	Test outcome is scoring above “proficiency” cutoff—greater weight on low achieving students	0.12 (0.07)	0.20 (0.14)	0.12 (0.08)	0.17 (0.13)
8	Estimate VA using only students <u>below class mean</u> prior achievement – use normalized gain score as outcome	0.21* (0.06)	0.37* (0.12)	0.29* (0.08)	0.44* (0.12)
9	Test outcome is percent correct squared—greater weight on high achieving students	0.20* (0.07)	0.29* (0.10)	0.26* (0.07)	0.33* (0.09)
10	Estimate VA using only students <u>above class mean</u> prior achievement – use normalized gain score as outcome	0.09 (0.07)	0.16 (0.15)	0.25* (0.09)	0.32* (0.11)
11	Test outcome is percentile in district test distribution	0.20* (0.07)	0.30* (0.10)	0.28* (0.07)	0.36* (0.09)
12	Outcome is the gain score normalized around predicted gain for students with comparable prior achievement	0.18* (0.07)	0.28* (0.10)	0.28* (0.08)	0.36* (0.10)
	<i>Alternative Specifications of the Value-Added Model</i>				
13	Exclude students in special education	.258* (.077)	.350* (.100)	.317* (.077)	.402* (.096)
14	Only use 2002-2003 achievement data	.101 (.093)	.191 (.267)	.191 (.111)	.233 (.135)
15	Measure value-added within grade-school, rather than simply within school	.201* (.063)	.428 (.309)	.228* (.081)	.321* (.150)
16	Gain outcome with student fixed effects	0.16* (0.07)	0.27† (0.14)	0.31* (0.08)	0.48* (0.16)
17	Include indicator for first-year teachers (+ polynomials in prior achievement)	0.19* (0.07)	0.30* (0.11)	0.24* (0.08)	0.30* (0.10)
18	Include ln(experience) variable (+ polynomials in prior achievement)	0.20* (0.06)	0.46 (0.35)	0.26* (0.07)	0.40* (0.12)
	<i>Other Checks</i>				

19	Block Bootstrapped Standard Errors (Clustering at School Level)	0.18* (0.09)	0.29* (0.14)	0.25* (0.09)	0.32* (0.12)
20	Average of principal-level correlations	0.21* (0.08)	--	0.31* (0.08)	--

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 VI  
HETEROGENEITY OF EFFECTS

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	0.18*	0.29*	0.25*	0.32*
		(0.07)	(0.10)	(0.08)	(0.10)
<b><i>By teacher characteristics</i></b>					
2	Experienced teachers ( $\geq 11$ years, n=92)	0.29*	0.35*	0.34*	0.39*
		(0.09)	(0.11)	(0.09)	(0.11)
3	Inexperienced teachers ( $< 11$ years, n=109)	0.07	0.38	0.13	0.20
		(0.08)	(0.45)	(0.11)	(0.19)
4	Principal known teacher for long time ( $\geq 4$ years, n=114)	0.22*	0.28*	0.29*	0.35*
		(0.09)	(0.11)	(0.08)	(0.10)
5	Principal hasn't known teacher for long ( $< 4$ years, n=87)	0.13	0.49	0.21	0.29
		(0.10)	(0.63)	(0.12)	(0.18)
6	Grades 2-4 (n=128)	0.30*	0.43*	0.36*	0.46*
		(0.07)	(0.11)	(0.07)	(0.09)
7	Grades 5-6 (n=73)	0.09	0.40	-0.10	-0.14
		(0.10)	(0.66)	(0.19)	(0.30)
<b><i>By principal characteristics</i></b>					
8	Principal has been in the same school at least 4 years (n=108)	0.08	0.12	0.12	0.16
		(0.10)	(0.15)	(0.10)	(0.16)
9	Principal has been in the same school less than 4 years (n=93)	0.29*	0.49*	0.35*	0.44*
		(0.09)	(0.25)	(0.10)	(0.13)
10	Principal confident in reading ratings (n=106)	0.21*	0.36*	0.37*	0.48*
		(0.08)	(0.16)	(0.09)	(0.12)
11	Principal not confident in reading ratings (n=79)	0.11	0.16	0.11	0.14
		(0.11)	(0.16)	(0.12)	(0.16)
12	Principal confident in math ratings (n=47)	0.32*	0.58	0.51*	0.82
		(0.11)	(0.49)	(0.11)	(0.82)
13	Principal not confident in math ratings (n=138)	0.12	0.17	0.17	0.21
		(0.08)	(0.12)	(0.10)	(0.13)
14	Male principal (n=103)	0.22*	0.32*	0.35*	0.46*
		(0.09)	(0.13)	(0.10)	(0.13)
15	Female principal (n=98)	0.16	0.25	0.19†	0.25†
		(0.10)	(0.16)	(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 VII  
THE ASSOCIATION BETWEEN DIFFERENT TEACHER QUALITY MEASURES AND FUTURE STUDENT ACHIEVEMENT

Independent Variables	Dependent Variable											
	2003 Reading Score						2003 Math Score					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
3-5 years of experience	-0.013 (0.061)						0.222 (0.125)					
6-10 years of experience	0.017 (0.060)						0.023 (0.125)					
11-20 years of experience	-0.019 (0.049)						0.127 (0.118)					
21+ years of experience	-0.049 (0.070)						0.009 (0.126)					
MA Degree	0.106* (0.046)						0.086 (0.075)					
Annual pay (in \$1,000)		0.000 (0.003)				-0.001 (0.002)		-0.000 (0.004)				0.000 (0.003)
Overall principal rating			0.070* (0.020)			0.045* (0.020)			0.141* (0.023)			0.077* (0.025)
Principal rating of ability to raise reading (math) scores				0.051* (0.019)						0.100* (0.029)		
Reading (math) value-added (EB measure)					0.106* (0.015)	0.096* (0.017)					0.207* (0.022)	0.176* (0.023)
R-squared	0.477	0.475	0.479	0.477	0.484	0.486	0.389	0.383	0.398	0.391	0.413	0.417

Notes: Each column represents a separate specification. Specifications in columns 1-6 include 162 teachers and 3,891 students; columns 7-12 include 118 teachers and 2,590 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.

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

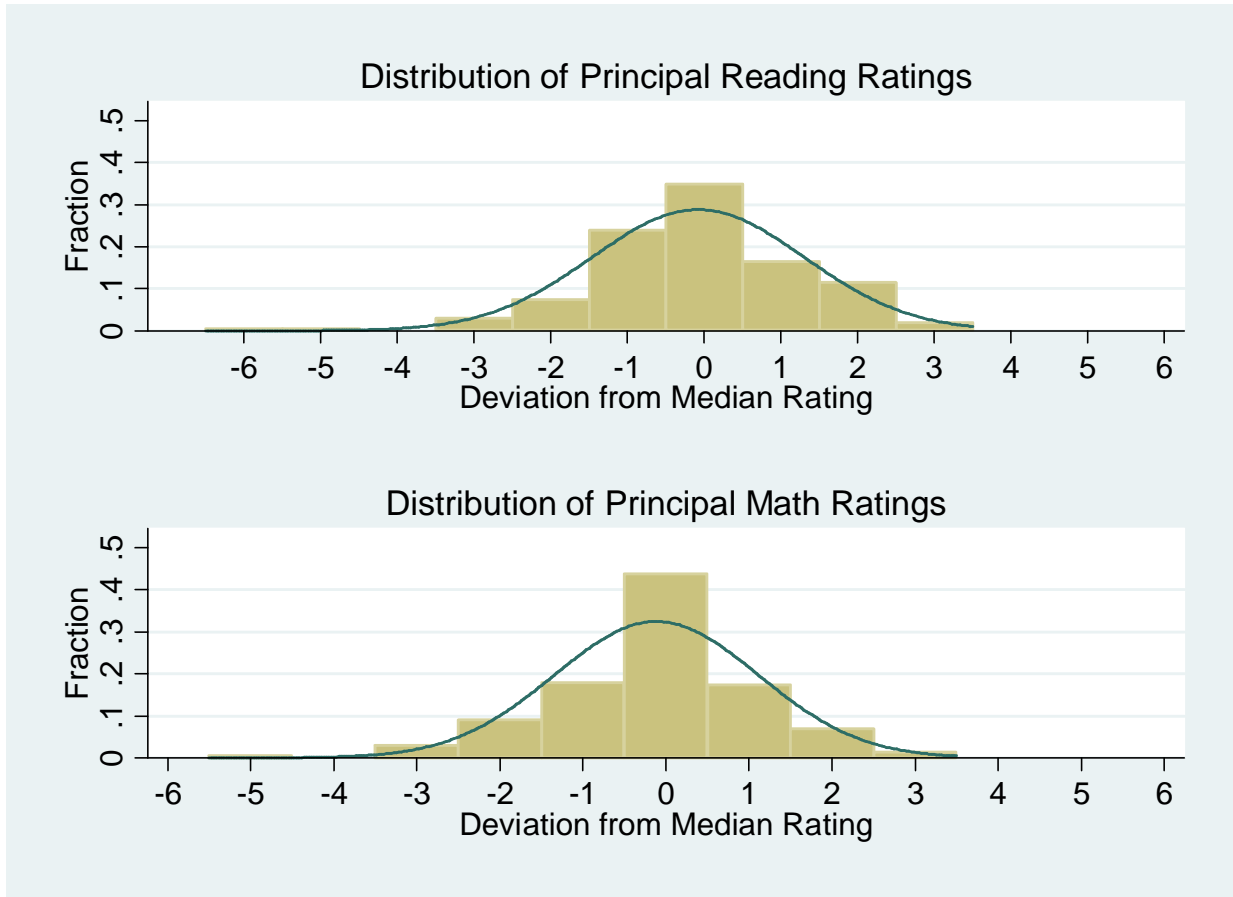


FIGURE II  
ASSOCIATION BETWEEN ESTIMATED TEACHER VALUE-ADDED AND PRINCIPAL RATING

