ASSESSING PRINCIPALS' ASSESSMENTS: SUBJECTIVE EVALUATIONS OF TEACHER EFFECTIVENESS IN LOW- AND HIGH-STAKES ENVIRONMENTS

Jason A. Grissom
(corresponding author)
Peabody College
Vanderbilt University
Nashville, TN 37203-5721
jason.grissom@vanderbilt.edu

Susanna Loeb
Center for Education Policy Analysis
Stanford University
Stanford, CA 94305-3084
sloeb@stanford.edu

Abstract

Teacher effectiveness varies substantially, yet principals' evaluations of teachers often fail to differentiate performance among teachers. We offer new evidence on principals' subjective evaluations of their teachers' effectiveness using two sources of data from a large, urban district: principals' high-stakes personnel evaluations of teachers, and their low-stakes assessments of a subsample of those teachers provided to the researchers. We find that principals' evaluations of teachers are quite positive whether the stakes are high or low, but the low-stakes evaluations show substantially more use of lower rating categories, and many teachers rated ineffective on the low-stakes assessment receive "effective" or "highly effective" high-stakes ratings. Teacher characteristics, such as experience, partially explain the discrepancy between the two scores. Also, despite the fact that principals overwhelmingly assign teachers to the two highest rating categories on the high-stakes evaluation, their high- and low-stakes ratings show similar correlations with teacher value-added measures.

doi:10.1162/EDFP_a_00210

1. INTRODUCTION

Education researchers and policy makers are currently engaged in a vigorous debate about how best to evaluate teachers. Much of this debate has centered on the use of student test score–based measures of teacher performance, or teacher "value added" (Glazerman et al. 2010; Darling-Hammond et al. 2012). Critics urge caution in relying on these measures in high-stakes personnel decisions (such as contract renewal) because of concerns about their reliability and validity (Rothstein 2010). Yet the current push to utilize value-added measures arose in response to evidence indicating that previous teacher evaluation systems based on principals' subjective performance assessments failed to differentiate among high and low performers. In many systems, virtually all teachers are given satisfactory ratings by their principals, even in schools with very low student achievement (Weisberg et al. 2009). These patterns have persisted as states have aimed to develop more rigorous evaluation systems in response to Race to the Top and other reforms (Sawchuk 2013). The quality of principals' subjective assessments of teachers remains an important policy concern in light of their use for high-stakes decisions about teachers and pressure to use teacher effectiveness data for decision making in schools more generally (Goldring et al. 2015).

It is unclear whether undifferentiated or inflated ratings of some teachers' performances on high-stakes evaluations result from a lack of capacity among principals to distinguish low-performing teachers or principals' unwillingness to give negative ratings even when they observe poor performance (Medley and Coker 1987; Frase and Streshly 1994). Several recent studies have—from surveys or interviews—compared principals' low-stakes assessments of teachers to teacher value-added scores, and found that principals give higher ratings to teachers with high value-added estimates, and lower ratings to teachers with low value added, which researchers have interpreted as evidence that principals can indeed identify their effective and ineffective teachers, especially in the tails of the distribution (e.g., Jacob and Lefgren 2008; Harris and Sass 2014). The context for these prior studies, however, is quite different from real-world evaluation in which teachers receive principals' ratings, have a feedback conversation about the ratings, and potentially face personnel action as a result of the ratings. Because of this contextual difference, we cannot necessarily expect principals to rate teachers on end-of-year evaluations in the same way they do when participating in research. In particular, we might expect that in low-stakes settings, principals may provide something closer to their "true" evaluation of the teacher's strengths and weaknesses, resulting in a larger correlation with other measures of teacher effectiveness (Epstein 1985; Harris, Ingle, and Rutledge 2014). In practice, political forces and managerial roadblocks, such as unwillingness to engage in difficult conversations, may distort principals' ratings, muting or even removing the relationship between high-stakes evaluation ratings and actual job performance. Some studies that have examined the relationships between student test score growth measures and principals' assessments of teachers in high-stakes settings have found positive correlations (e.g., Sartain, Stoelinga, and Brown 2011), but changes in test scores likely only capture one area of teacher job performance, so this comparison may not shed much light on the question of how evaluation ratings reflect principals' actual perceptions of teacher performance across the range of job activities schools value. Research has not yet closed the loop among these related strands of research by comparing principals' low-stakes and

high-stakes assessments of teachers and considering those assessments' relative predictive validity with respect to value-added measures.

To address this gap, this study investigates principals' evaluations of teachers in a large, urban school district. We draw on two sources of teacher evaluation data. First, we use principals' summative evaluations of teachers on the school district's formal personnel evaluation instrument for the 2011–12 academic year. Principals' subjective assessments on this instrument are "high stakes" in the sense they inform personnel decisions, such as the nonrenewal of low-performing teachers. Second, we utilize data from a subset of principals evaluating teachers during structured interviews in approximately 100 schools in the district during the spring of 2012. During these interviews, we provided principals with the names of three to four teachers in their school and asked them to rate those teachers on a six-point scale in eight areas of job performance. We also asked them to provide overall ratings of the teachers' instruction and performance in noninstructional areas. We match principals' evaluations from these interviews to the district evaluation they conducted at the end of the school year, allowing us to compare ratings of the same teachers by the same principals in low- and high-stakes environments. We use longitudinal administrative personnel and student data provided by the district, including student test scores, to link the principals' evaluations to estimates of teachers' value added to student test performance.

The construction of this unique dataset allows us to answer the following research questions. First, what is the distribution of principals' teacher ratings, and does this distribution vary by whether the evaluation is high or low stakes? Second, how are teachers' scores on the high- and low-stakes assessments compared with one another and with value added? That is, do teachers who do well on one performance measure also do well on the others, and do principals' assessments of teacher performance predict value added differently when different stakes are attached? And finally, what principal and teacher characteristics predict whether a teacher's official personnel evaluation score is higher or lower than what would be predicted by the low-stakes rating?

Our results show that principals' ratings of teachers in both the high- and low-stakes evaluations are negatively skewed. Principals are, however, much more likely to give low ratings on the low-stakes assessment. On the high-stakes instrument, the distribution is truncated, with nearly all teachers receiving scores of "effective" or "very effective" on every standard. In fact, fewer than 3 percent of teachers in the district received a score less than "effective" on *any* of the seven standards in 2012. Scores across items within instruments are moderately to highly positively correlated. In fact, factor analysis of each sets of evaluations reveals a single latent construct, which we interpret as perceived job performance. We find that although scores on the high- and low-stakes evaluations are moderately correlated with one another, these correlations mask important differences in the absolute ratings principals give on the two instruments—even teachers given scores of "very ineffective" on classroom effectiveness on the low-stakes instrument receive scores averaging above 3.0 ("effective") on the high-stakes instrument. Yet despite the truncation in the distribution in the high-stakes assessment, the two evaluation instruments are generally similarly correlated with teacher value added, though more so for math than reading.

When we examine the difference between *actual* high-stakes evaluation scores and the high-stakes scores we would have predicted based on the low-stakes rating, we find

that some teachers are evaluated better than predicted. For example, novice teachers score higher than their low-stakes evaluations would predict. We also find evidence of principal- or school-level idiosyncrasies in the propensity to rate teachers high or low, but observed school and principal characteristics explain little of this variation. Overall, our findings are consistent with studies suggesting that principals can identify which teachers are high performers. At the same time, they illuminate potential limitations of high-stakes teacher evaluations, given principals' apparent reluctance to identify teachers as ineffective when ratings matter, even when they have done so on low-stakes evaluations.

2. PRINCIPAL ASSESSMENTS OF TEACHER EFFECTIVENESS

State policy changes around teacher evaluation in response to Race to the Top and the No Child Left Behind waiver process, as well as recent investments in measuring teacher effectiveness by foundations (such as the Bill and Melinda Gates Foundation's Empowering Effective Teachers initiative), show a renewed interest in improving the quality of teacher evaluation nationwide. Recent policy changes typically structure evaluation systems around multiple measures of teacher performance, which most often include measures constructed from student test scores and derived from principals' observations of teacher instruction, although other measures, such as student or parent perceptions surveys or peer observations, are growing in prevalence. Even though the student achievement or growth components of these systems have generated a great deal of scholarly, policy, and media attention, the fact that observation-based measures often constitute half of the overall evaluation score—and the fact that many districts are using these evaluation scores for high-stakes personnel decisions, such as compensation and dismissal—highlights the importance of the principal's role (Doherty and Jacobs 2013; Drake et al. 2016). An assumption underlying these systems is that principals are capable and willing to accurately and reliably evaluate teacher performance. Relatively little systematic attention has been paid to how equipped principals are to serve in this role, however.

Several recent studies have examined principals' ratings of teacher effectiveness in low-stakes settings, all pointing to the principals' ability to distinguish teachers with high and low value added to student test performance. Jacob and Lefgren (2008) linked principals' subjective assessments of teachers' skills, collected via a survey of elementary school principals in an anonymous midsize district in the western United States, to the test performance of the teachers' students. The researchers asked principals to rate teachers in their schools ($N = 202$ teachers) on a scale of 1 to 10 measuring "overall teacher effectiveness" and, more specifically, for their skills at raising student achievement in math and (separately) reading. They found that principals showed some adeptness at identifying which teachers were more able to raise student achievement in math and reading, with ratings on those factors modestly predicting teachers' value added in those subjects across a variety of modeling specifications (correlations are generally about 0.30). Principals were more able to identify the highest- and lowest-performing teachers in their subjective ratings but showed less capacity to differentiate teachers near the middle of the effectiveness distribution, as measured by value added.

Similarly, in the context of an experimental evaluation of a New York City initiative to provide principals with value added information about their teachers, Rockoff et al. (2012) report that principals' overall subjective evaluations of teachers (at baseline), gathered via surveys, correlated to teachers' value added in math and reading at approximately 0.25. They also found these correlations are higher when value-added estimates are more precise and for more experienced principals.

Two other studies examined interview data from thirty principals in an anonymous midsized district in Florida collected during 2005 and 2006 to address a related set of issues (Harris, Ingle, and Rutledge 2014; Harris and Sass 2014). Researchers asked each principal to rate up to ten teachers of tested grades and subjects, resulting in a sample size of approximately 235 teachers. Principals provided both overall subjective ratings of teacher effectiveness on a 1 to 9 scale and ratings of teachers on eleven job traits, including caring, knowledge of subject, strong teaching skills, motivation, and contributions to school activities beyond the classroom. They then reduced these items to four dimensions via factor analysis, which Harris and Sass (2014) labeled interpersonal skills, motivation/enthusiasm, ability to work with others, and knowledge/teaching skills/intelligence.[1] The authors found that principals' overall assessments of teacher performance generally correlated with value added in math and reading in the range of 0.30 across all schools. Among the latent trait measures, the knowledge trait was most highly correlated with math value added, and the motivation trait was most highly correlated with reading value added.[2] In an additional analysis of qualitative responses provided by principals describing their ratings of teachers, Harris, Ingle, and Rutledge (2014) uncovered evidence that teacher personality, philosophy, and effort contributed to principals' ratings in ways that explained the divergence between ratings and value-added scores.

These studies examine the relationship between low-stakes research-based assessments and value added; another set has focused instead on the associations between student test score growth and principals' ratings on high-stakes instruments mandated by school districts. For example, Kimball et al. (2004) examined data from the implementation of the Framework for Teaching (FFT) as an evaluation instrument in Washoe County, Nevada—scores from which had potential consequences for teachers (such as referral to an intervention process or initiation of a dismissal process). Principals' evaluations of teachers were only statistically significantly associated with student achievement growth in half of the grade/subject combinations examined. More recently, in a study of a two-year pilot of an evidence-based teacher evaluation system in Chicago, Sartain, Stoelinga, and Brown (2011, also using FFT) found that principals' ratings of teachers were strong predictors of value added in both math and reading. Across FFT components related to classroom environment (domain 2) and instruction (domain 3), differences in value added between teachers rated *unsatisfactory* and *distinguished* ranged from 0.3 (component 2a) to 0.9 (component 2b) standard deviation for reading and 0.4 (component 2a) to 1.0 (component 3a) standard deviation for math. The study

1. Harris, Ingle, and Rutledge (2014) similarly arrive at four factors, though they provide a different labeling.
2. Harris, Ingle, and Rutledge (2014) find that a factor they call "technical skill" is the best predictor of both math and reading growth.

also finds that principals were much more likely to give the highest rating to teachers than were external observers who simultaneously scored the same lesson.

Our study is the first in this line of research to examine principals' teacher evaluations using both high- and low-stakes assessment information collected from the same principal. Prior research has effectively demonstrated that principals are capable of identifying effective teachers in their schools in low-stakes settings, and research also shows that ratings of teachers using rigorous observation instruments in some contexts correlate with teacher value added and other effectiveness measures. We build on this work by comparing principals' low-stakes and high-stakes evaluations of the same teachers, examining their patterns of convergence and divergence, and analyzing their relative magnitudes of association with value-added effectiveness measures.

Why is this analysis important? Districts combine principal assessments of teachers with value-added measures for a reason. Value-added measures likely pick up some dimensions of teacher performance, but not all dimensions. They also are subject to idiosyncratic error that may come from unpredicted shocks in classrooms, such as conflict among students unrelated to teacher performance. Principals can pick up on these idiosyncratic effects as well as assess teachers on factors not measured well by value added. On confidential low-stakes evaluations, principals have little incentive not to provide their true assessment of teachers. However, on high-stakes assessments they might have such incentives (e.g., MacLeod 2003). By comparing the two, we can see whether principals change their reports given this difference in context. By comparing each to value added, we can also potentially see whether such change indicates movement away from principals' assessment of teachers—which reflects their value added—versus principals' assessment of teachers that shows either the idiosyncratic factors or the dimensions of good teaching that are not well captured by value added.

3. DATA AND MEASURES

We examine principals' subjective evaluations of teachers using data from an ongoing study of school leadership in Miami-Dade County Public Schools (M-DCPS) (e.g., Grissom and Loeb 2011; Grissom, Loeb, and Master 2013). M-DCPS is the largest public school district in Florida and the fourth largest in the United States, enrolling approximately 350,000 students across close to 400 schools. Nearly 90 percent of students in the district are either black or Hispanic, and 60 percent qualify for free or reduced-priced lunches. We combine formal evaluation scores with original data describing students, teachers, and schools that we collected from interviews with principals and administrative records provided to us by the district.

High-Stakes Evaluation Ratings

In 2012, the focal year for this study, M-DCPS required the evaluation of all instructional personnel using the district's educator assessment tool, the Instructional Performance Evaluation and Growth System, or IPEGS. Teachers' IPEGS evaluations require a formal summative evaluation by the principal that uses the IPEGS tool at the end of the school year.

Table 1. Names of Standard, Percent of Teacher Evaluation Score, and Descriptive Statistics for High-Stakes Scores

Standard	Percent of Evaluation Score	Obs.	Mean	Std. Dev.	Min	Max
2 Knowledge of Learners	8	22,402	3.58	0.51	1	4
3 Instructional Planning	8	22,402	3.52	0.51	1	4
4 Instructional Delivery and Engagement	8	22,392	3.52	0.52	1	4
5 Assessment	6	22,394	3.40	0.50	1	4
6 Communication	6	22,389	3.55	0.51	1	4
7 Professionalism	6	22,400	3.50	0.53	1	4
8 Learning Environment	8	21,154	3.57	0.51	1	4

Notes: Standard 1 is Learner Progress, which depends on student test score growth and constitutes the remaining 50 percent of a teacher's evaluation rating. Scores converted to 4-point scale with 1 = *unsatisfactory*, 2 = *developing/needs improvement*, 3 = *effective*, and 4 = *highly effective*.

IPEGS rates teachers on eight standards. On each standard, teachers are scored as *highly effective, effective, developing/needs improvement,* or *unsatisfactory*. Standard 1, which is worth 50 percent of the overall evaluation score, is Learner Progress; scores on this standard are calculated by the district using student test score growth measures. The remaining seven standards make up the other 50 percent of a teacher's score and are rated subjectively by the principal using a matrix that describes what teacher performance looks like on that standard at each of the four performance levels. Table 1 shows the names of the standards along with what percentage of the overall evaluation score each standard contributes (ranging from 6 percent to 8 percent) and some descriptive statistics about teachers' scores. M-DCPS provided us with IPEGS ratings on each standard for all of the district's instructional staff for the 2011–12 school year.

Low-Stakes Ratings in Principal Interviews

In the spring of 2012, we conducted structured interviews with 98 M-DCPS principals. The interview sample included nearly all high school principals ($N = 41$) in the district, plus a random sample of twenty-eight elementary/K–8 and twenty-nine middle school principals.[3] For one portion of the interview, we asked principals to discuss the strengths and weaknesses of four different teachers whose names the interviewer provided. Each of the teachers was an instructor of a tested grade and subject combination for whom we could calculate value added in math and/or reading as of the prior school year.[4] To ensure variation, we chose two teachers at random from below the median of the 2010–11 value-added distribution and two from above the median. Whenever possible, we chose one relatively inexperienced teacher (i.e., less than five years) and one experienced teacher (five or more years) from below and above the value-added median.

We asked principals a series of open- and closed-ended questions about each teacher, first about that teacher's performance with students in the classroom and then

3. We initially aimed to collect data in forty-five high schools, thirty elementary/K–8 schools, and thirty middle schools, but were unable to complete all school visits because of scheduling conflicts with school personnel.
4. Because the 2011–12 school year had not yet ended at the time of the interviews, we could not know for sure that every teacher in the sample would have sufficient data for calculating a value-added score for that year. A small number of teachers dropped out of the analysis because they ultimately could not be matched to a 2011–12 value-added score.

Table 2. Descriptive Statistics for Interview Ratings

	Mean	SD	Min	Max	*N*
"In-Class" Items					
Getting High Test Performance	4.64	1.16	1	6	355
Improving Critical Thinking	4.56	1.28	1	6	371
Motivation	4.75	1.24	1	6	372
Building Interpersonal Skills	4.68	1.20	1	6	373
Overall In-Class Effectiveness	4.77	1.14	1	6	374
"Out-of-Class" Items					
Building Staff Relationships	4.40	1.34	1	6	371
Supporting Colleague Instruction	4.40	1.39	1	6	371
Helping with Leadership/Management	4.31	1.43	1	6	362
Building Community Relationships	4.14	1.45	1	6	365
Overall Out-of-Class Effectiveness	4.34	1.31	1	6	370

about his or her performance in areas outside the classroom.[5] For in-class items, we first asked, "What are this teacher's strengths and weaknesses inside the classroom, that is, in his/her role as an instructor?" We then asked about four specific areas of instruction, asking the principal to rate the teacher on a scale of 1 (very ineffective) to 6 (very effective): (1) getting high standardized test performance from students; (2) developing students' higher-order thinking skills, such as synthesis and evaluation; (3) motivating students to learn; and (4) helping students build strong interpersonal skills. We also asked the principal to rate the teacher's overall classroom performance using the same 1 to 6 scale.[6]

Next, we asked principals to discuss the teachers' contributions outside the classroom: "I'd like you to think about this teacher's strengths and weaknesses outside his/her own classroom. In other words, I'd like to know about how this teacher contributes (or doesn't contribute) to the school environment beyond his/her role as a classroom instructor. How would you describe these strengths and weaknesses?" We then again asked the principal to rate each teacher on four nonclassroom performance dimensions using a 1 to 6 scale: (1) building positive interpersonal relationships among the staff at this school; (2) supporting the instructional effectiveness of his/her fellow teachers (e.g., through mentoring or being a resource); (3) helping you and your leadership team manage the school effectively (e.g., by taking on leadership roles or being "someone you can call on"); and (4) building a strong relationship between the school and the community outside the school, including parents. We also asked the principals to rate the teachers' overall performance outside the classroom.

Table 2 describes the principals' ratings of teachers. In terms of rank order, principals rated teachers highest on their ability to motivate students, and lowest in their role in building community relations. Ratings for overall

5. We do not make use of responses to the open-ended items in this analysis.
6. The specific question was: "Now thinking about all aspects of this teacher's classroom performance, not just the ones I've already mentioned, how would you rate this person's overall effectiveness as a teacher using the 1 to 6 scale?"

instructional/in-classroom effectiveness were 4.8, on average, on a six-point scale, while ratings for noninstructional/out-of-classroom effectiveness averaged 4.3.

Administrative Data

M-DCPS provided us with three longitudinal administrative data files: background information for all students in the district, course-level data that link students to each of their teachers in a year, and a staff file with information on all district employees, including the school in which they work. Information on the students includes race and ethnicity, gender, eligibility for free or reduced-priced lunch, attendance, suspensions, and test scores from the state's standardized testing program; tests in reading and math are administered to students in grades 3 through 10. Staff data include the highest degree earned, years of experience in the district for teachers, years in the job for principals, race, ethnicity, gender, and number of work absences. These data span the 2003–04 through the 2011–12 school years and are used both in creating value-added measures for teachers and as covariates in some analyses.

Estimating Teacher Value Added

We estimated six measures of value added for each teacher when possible, three each for math and reading. First, we calculated average value added over all available years of data. Equation 1 describes this model:

$$A_{itgsy} = \beta_1 A_{itgs(y-1)} + \beta_2 A^{Other}_{itgs(y-1)} + X_{itgsy}\beta_3 + C_{tgy}\beta_4 + S_{sy}\beta_5 + \pi_g + \mu_y + \delta_t + \varepsilon_{itgsy}. \quad (1)$$

Achievement A for student i with teacher t in grade g in school s in year y is a function of the student's prior test performance both in the same subject and in the other subject, student characteristics X, classroom characteristics C, school characteristics S, and grade, year, and teacher fixed effects. The parameter δ, the teacher fixed effect, reflects the contribution of a given teacher to student achievement after controlling for all observed student, classroom, and school characteristics, over all available years of data. The test scores used to generate the value-added estimates in each subject are scale scores, standardized to have a mean of zero and a standard deviation of one for each grade in each year.

Next, we estimated value added in the 2011–12 school year only. For this estimate, we essentially re-estimate equation 1 except that instead of including a teacher effect and a year effect, we include a teacher-by-year fixed effect. The coefficients on the teacher-by-year fixed effects corresponding to 2011–12 are then used as estimates of a teacher's impact on student achievement in math or reading in that year.[7]

Lastly, using a model developed and applied by Chetty, Friedman, and Rockoff (2014), we create value-added estimates that account for drift in teachers' value added to student achievement across school years. Rather than assume that a teacher's "quality" is fixed, and thus equally weighting all available years of data, the Chetty, Friedman,

7. The estimated coefficients for these fixed effects include measurement error as well as real differences in achievement gains associated with teachers or schools. Thus, in some analyses we shrink the estimates using the empirical Bayes method to bring imprecise estimates closer to the mean (see details in Appendix A).

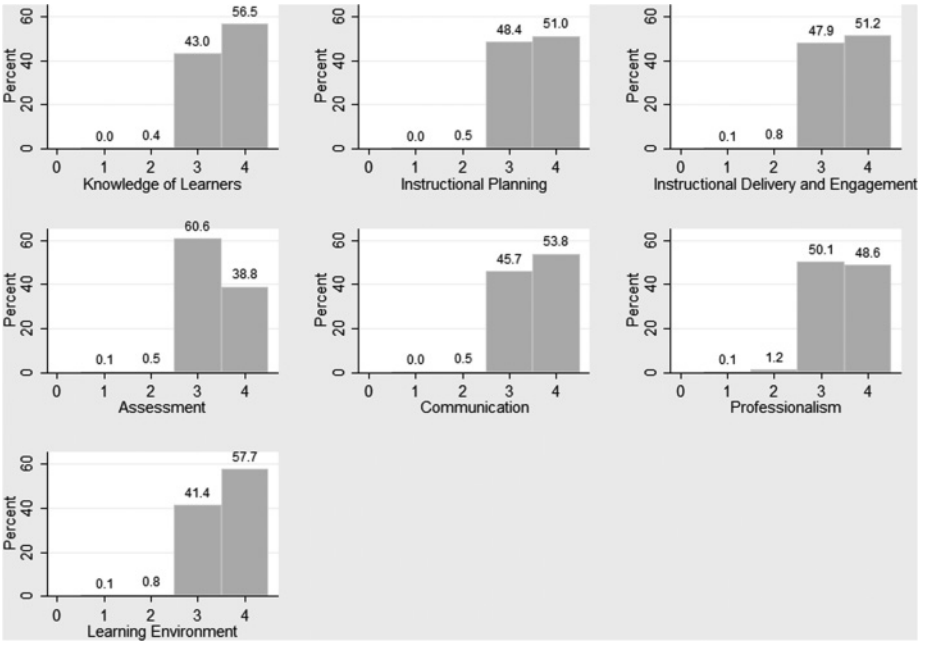


Figure 1. Distribution of Scores on High-Stakes Evaluation Instrument.

and Rockoff (2014) method explicitly allows the value added of a teacher to change over time so that an estimate for 2011–12, for example, puts greater weight on more recent years of data than on years earlier in the dataset. (Further details and Stata programs for implementing the Chetty, Friedman, and Rockoff approach are available at www.rajchetty.com/chettyfiles/value_added.htm.)

Note that the fraction of teachers for whom value added can be calculated in math or reading ranges from 18 to 31 percent of the full teacher sample, depending on the measure type. Appendix table A.1 shows correlations among the measures. For teachers with both math and reading scores, average math and reading value added correlates at 0.70. The 2011–12 values correlate at 0.58, and the drift-adjusted values correlate at 0.86. In math, the average and 2011–12 value-added estimates correlate at 0.62; in reading, they correlate at 0.57. In both math and reading, the drift-adjusted measure correlates more highly with the average measure (0.62 and 0.41, respectively) than with the 2011–12 measure (0.46 and 0.24, respectively).

4. RESULTS

Our analyses of the relationships among principals' high- and low-stakes evaluations of teachers and teachers' value-added scores address several research questions. We present results below, organized by question.

Question 1: How are principals' ratings of teachers distributed, and do these distributions vary by whether the evaluation is high- or low-stakes?

On both sets of evaluations, principals assess most teachers as effective across job dimensions. Figure 1 shows the distribution of scores, with a possible range of 1 to 4, on

each standard on the high-stakes performance assessment instrument. The scores for each standard are lumpy and very negatively skewed. Principals almost never assign teachers scores of 1 or 2 on any standard (fewer than 1 percent for any standard except Professionalism). In fact, only 566 teachers, or 2.9 percent of teachers with ratings, obtained a score of 1 or 2 on *any* standard. Nevertheless, there is some variation for each standard, with relatively similar numbers of teachers receiving scores of *effective* (3) and *highly effective* (4) for each one. For example, 48 percent of teachers received a score of *effective* for Instructional Delivery and Engagement, and 51 percent of teachers received a score of *highly effective*.

Figure 2 shows the distribution of scores from the low-stakes interviews. Again, the scores are negatively skewed, although principals are more likely to assign teachers scores in the lower performance rating categories on the low-stakes assessment. Across the four in-class items, the fraction of teachers assigned a score of 3 or below (out of 6), indicating *a little* to *very ineffective*, ranged from 17 to 21 percent, with nontrivial percentages given a score of 1 or 2. For example, among the 21 percent of teachers rated 3 or below for improving the critical thinking of their students, 43 percent (9 percent of total) were given a score of 1 or 2. Teachers were given low ratings even more frequently for the out-of-class dimensions, with the fraction rated 3 or below ranging from 26 percent for building relationships with fellow staff members (10 percent received a 1 or a 2) to 34 percent for building relationships with the community (14 percent received a 1 or a 2). For the two summative items, teacher ratings were similarly negatively skewed. For overall in-class effectiveness, 15 percent were rated 3 or below, with 4 percent receiving a 1 or a 2, and for overall out-of-class effectiveness, 29 percent were rated 3 or below, with 9 percent scoring 1 or 2.

Question 2: How do teachers' scores on the high- and low-stakes assessments compare with one another and with value-added?

Next, we assess how principals' ratings of teachers compare across the high- and low-stakes assessments. A Spearman rank correlation matrix among the items within the two rating sets (shown in Appendix table A.2) shows moderate correlations within the high-stakes evaluation ratings, ranging from 0.36 to 0.53, and somewhat higher correlations among the low-stakes interview ratings, particularly within the in-class or out-of-class item sets, ranging from 0.47 to 0.87. We then conducted separate exploratory factor analyses on the two sets of evaluation ratings.[8] In both cases, one underlying construct was clearly identified based on scree plots of the eigenvalues. For the high-stakes evaluation, the items making up this single factor had a Cronbach's α of 0.84. For the low-stakes evaluation items, Cronbach's $\alpha = 0.94$. These results suggest that principals do not differentiate teacher performance across the various dimensions on either instrument—either because differentiating is difficult or because the dimensions are indeed highly correlated within teachers—but instead have a single underlying perception of job performance for each teacher that dictates the principal's ratings across the items.

8. Because ratings fall into ordinal categories, we based each factor analysis on polychoric correlation matrices. For the low-stakes ratings, we conducted factor analysis on items including and excluding the "overall in-class" and "overall out-of-class" variables and found them to be very similar; factor scores were correlated at 0.99. Here we report results that include the items.

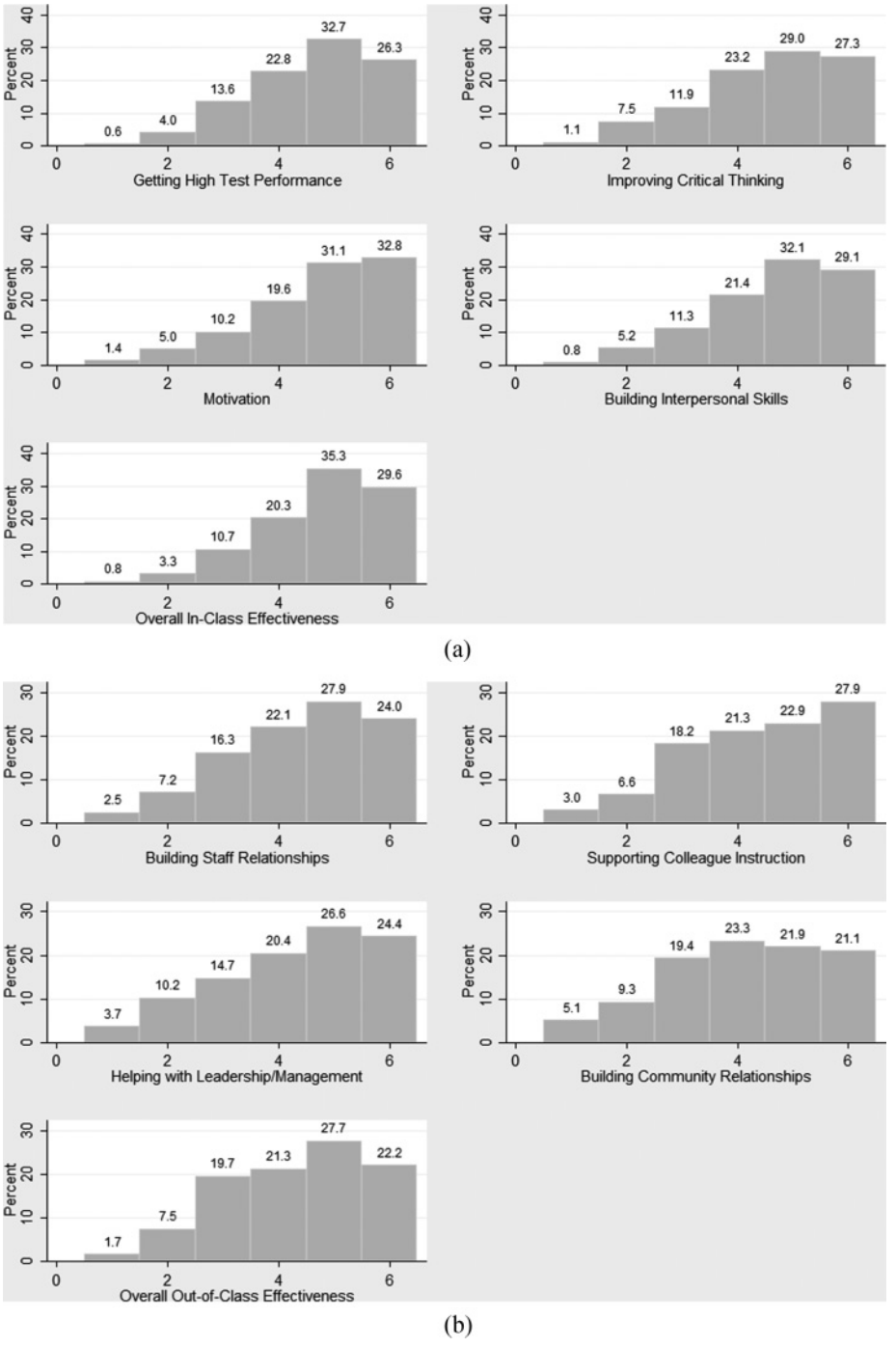


Figure 2. Distribution of Scores from Low-Stakes Interview Ratings. a. Distribution of Teacher Ratings on In-Class Items. b. Distribution of Teacher Ratings on Out-of-Class Items.

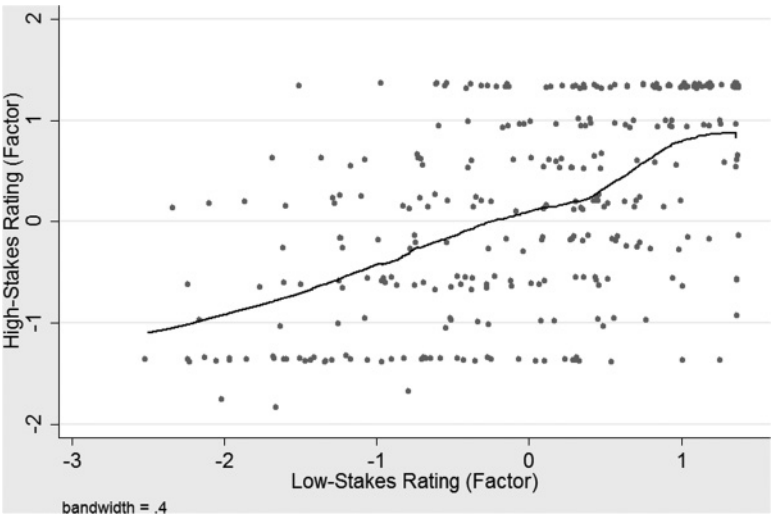


Figure 3. Locally Weighted Regression of High-Stakes Rating Factor on Low-Stakes Rating Factor.

The factor scores for the respective high- and low-stakes latent constructs correlate at 0.55.[9] Using the reliability measures to adjust for random error in the two factor scores, the correlation between the two underlying factors is 0.62.[10] Figure 3 illustrates the relationship between the two measures with a scatterplot and a nonparametric locally weighted regression line fit to the plotted points (i.e., Lowess curve). It shows that the association generally appears linear. This high correlation provides support for the idea that both evaluations identify a similar underlying job performance construct, despite differences between items measured by the two instruments. Histograms of the factor scores are shown in Appendix figures A.1 and A.2.

Table 3 shows how the high-stakes and low-stakes ratings correlate with different measures of value added in math and reading.[11] Correlations are adjusted for estimation error in the value-added assessments, as described in Jacob and Lefgren (2008, p. 113).[12] Panel A displays the correlations for the high-stakes evaluation from IPEGS. Within subject, correlations are similar by standard across the different value-added

9. Appendix table A.3 shows Spearman correlations for the rating items across the two evaluation types. Scores on the high- and low-stakes evaluation items are moderately correlated. For example, the correlation between the average across the seven standards (last column) on the personnel evaluation and the in-class effectiveness score from the interviews is 0.55; the correlation with out-of-class effectiveness is 0.50.
10. In classical test theory, this correction divides the observed correlation by the square root of the product of the two reliabilities (Gulliksen 1987). Application of this formula suggests that if the low- and high-stakes ratings both measured the exact same construct, given these reliabilities, the correlation between the two sets of scores would be 0.89.
11. Comparing teacher ratings to value added across principals may be problematic if principals vary in degree of leniency or how they scale ratings. Prior studies address this issue by normalizing principals' ratings by subtracting the school-specific mean for a particular rating category and dividing by the school's standard deviation for that rating (e.g., Jacob and Lefgren 2008). Because in this study we have low-stakes ratings for at most four principals in a school, we chose not to normalize ratings in the results presented in the paper. In analysis available upon request, however, we normalized both the low- and high-stakes ratings and found only minor differences in the correlations with the math and reading value-added measures.
12. In particular, we apply a correction calculated as the square root of the variance ($\hat{\delta}$) of the observed value-added measures divided by the square root of $\hat{\delta} - Var(e)$, where $Var(e)$ is the square of the standard errors of the value-added estimates.

Table 3. Correlations between Ratings Variables and Value-Added Scores

Panel A: High-Stakes Evaluation Instrument						
	Average Value Added, All Years		Value Added, 2011–12		Drift-Adjusted Value Added, 2011–12	
Standard	Math	Reading	Math	Reading	Math	Reading
2 Knowledge of Learners	0.19	0.15	0.16	0.02	0.20	0.12
3 Instructional Planning	0.17	0.13	0.17	0.07	0.18	0.13
4 Instructional Delivery and Engagement	0.23	0.18	0.26	0.14	0.26	0.19
5 Assessment	0.20	0.17	0.20	0.07	0.22	0.16
6 Communication	0.07	0.11	0.10	0.05	0.11	0.10
7 Professionalism	0.15	0.12	0.15	0.10	0.18	0.14
8 Learning Environment	0.15	0.15	0.18	0.07	0.23	0.17
High-Stakes Rating (Factor)	0.26	0.22	0.27	0.11	0.31	0.22
Panel B: Low-Stakes Interview Instrument						
	Average Value Added, All Years		Value Added, 2011–12		Drift-Adjusted Value Added, 2011–12	
Standard	Math	Reading	Math	Reading	Math	Reading
"In-Class" Items						
Getting High Test Performance	0.48	0.38	0.44	0.14	0.58	0.28
Improving Critical Thinking	0.41	0.23	0.36	0.17	0.50	0.30
Motivation	0.25	0.20	0.14	0.11	0.34	0.26
Building Interpersonal Skills	0.22	0.13	0.01	0.00	0.27	0.20
Overall In-Class Effectiveness	0.39	0.27	0.33	0.07	0.43	0.24
"Out-of-Class" Items						
Building Staff Relationships	0.18	0.00	0.04	0.16	0.19	0.09
Supporting Colleague Instruction	0.23	0.17	0.18	0.16	0.36	0.17
Helping with School Leadership/Management	0.14	0.21	0.07	0.26	0.29	0.09
Building Community Relationships	0.10	0.16	−0.09	0.07	0.22	0.07
Overall Out-of-Class Effectiveness	0.20	0.18	0.08	0.16	0.30	0.08
Low-Stakes Rating (Factor)	0.31	0.22	0.18	0.13	0.42	0.22

Notes: Correlations between the rating items and average and 2011–12 value-added measures are adjusted for measurement error in the value-added scores. Correlations with the factor scores are adjusted for measurement error in both the value-added scores and the factor scores.

measures, though they tend to be slightly higher for the drift-adjusted measures and higher for math than reading. Perhaps unsurprisingly, the highest correlations are for the Instructional Delivery and Engagement standard. Because the factor analysis of the ratings suggests that principals do not systematically differentiate performance on one item from another, the factor rating score shown at the bottom of the panel provides a useful summary of the rating instrument. It similarly shows systematically higher correlations for math and equal or higher correlations for the drift-adjusted measure.[13]

These patterns continue in panel B, which displays the correlations for the low-stakes interview ratings. The drift-adjusted measures are again generally more highly correlated with the items, and math value added is more highly correlated in most

13. Correlations with the factor scores are adjusted for measurement error in the factor score as well as error in the value-added measure.

Table 4. Correlations between Rating Factor Scores and Drift-Adjusted Value-Added Scores

	Math			Reading		
	Elementary/K–8	Middle	High	Elementary/K–8	Middle	High
High-Stakes Rating (Factor)	0.24	0.27	0.27	0.19	0.16	0.03
Low-Stakes Rating (Factor)	0.40	0.33	0.32	0.24	0.02	0.07

Notes: Drift-adjusted value-added measures (2011–12) used. Correlations are adjusted for measurement error in the factor scores.

cases than is reading. The in-class items are more highly correlated with value added than are the out-of-class items, with Getting High Test Performance and Improving Critical Thinking demonstrating the strongest associations. The low-stakes rating factor is more highly correlated with average and drift-adjusted math value added than is the high-stakes factor, but not the 2011–12 measure. The two rating variables are similarly correlated with value added in reading across measure types.

Table 4 breaks down the correlations by school level to consider the possibility that, despite using a common rubric, the criteria that a principal might use to evaluate a teacher might be quite different in elementary, middle, and high schools. For simplicity, we focus on the factor variables and show only the drift-adjusted measures, which were the most highly correlated with the ratings in table 3. There are some differences. For the high-stakes evaluation, principals' ratings are similarly correlated with value added in math across school levels. In reading, however, high school principals' ratings are uncorrelated with value added (elementary and middle schools are similar to one another). For the low-stakes interview rating, principals' assessments are more highly correlated with value added in both math and reading. Middle and high school principals' low-stakes ratings are essentially uncorrelated with value added in reading. Comparing the high- and low-stakes rating, it appears that the low-stakes rating generally is consistently more strongly correlated with value added across levels in math—most clearly for elementary and K–8 schools—but not for reading.

Figure 4 provides further illustration of the similar correlation with value added across principals' low- and high-stakes assessments of teachers. Each graph shows a scatterplot with a line fitted using locally weighted regression of drift-adjusted value added on the rating factor variables. The top panel shows math, and the bottom shows reading, with the left column showing low-stakes ratings and the right column showing high-stakes ratings. There are two observations. First, there are no clear nonlinearities in any of the four fitted lines. Second, the slopes of the lines are similar.[14] In other words, the information contained in the principals' ratings about teacher performance, at least as measured by value added, generally is similar regardless of the stakes attached.

Nevertheless, as noted previously, there are important differences in the *absolute* ratings principals give on the two instruments. As an illustration, in table 5 we show average scores on the high-stakes assessment broken down by scores on the total in-class effectiveness and total out-of-class effectiveness rating from the low-stakes assessment. Even teachers who receive scores of *very ineffective* on in-class effectiveness on the

14. These observations are robust to experimentation with other reasonable bandwidths for the Lowess curve.

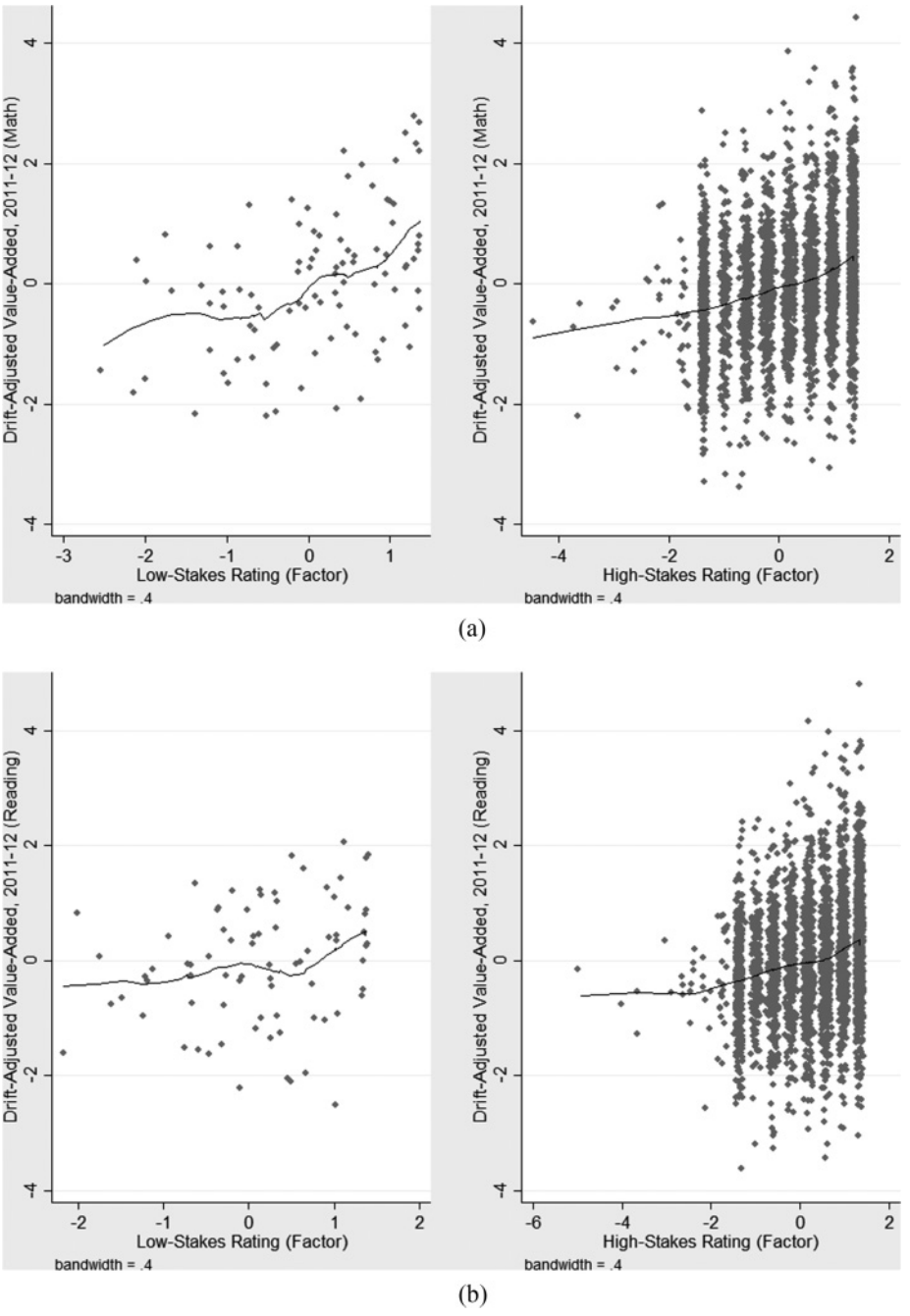


Figure 4. Locally Weighted Regression of Drift-Adjusted Value-Added Scores on Rating Factors. a. Math. b. Reading.

low-stakes instrument average above 3.0 (*effective*) on the high-stakes instrument. A similar pattern is observed for out-of-class effectiveness. Moreover, among the 13 teachers given either a 1 or a 2 on both the overall in- and out-of-class low-stakes ratings, the mean high-stakes score (averaging across the 7 standards) was 3.2, or slightly

Table 5. High Stakes Ratings and Value-Added Scores by Levels of Low-Stakes Ratings

	N	Average High-Stakes Score	*N*	Average Value Added, All Years	*N*	Value Added, 2011–12	*N*	Drift-Adjusted Value Added, 2011–12
Total In-Class Effectiveness								
1 (very ineffective)	2	3.14	0		0		0	
2	12	3.23	4	−0.44	4	−0.66	6	−0.99
3	39	3.20	15	0.05	15	−0.11	12	−0.42
4	71	3.37	27	0.08	25	−0.12	31	−0.57
5	120	3.60	45	0.17	45	0.07	52	−0.08
6 (very effective)	94	3.77	52	0.61	47	0.30	56	0.48
Total Out-of-Class Effectiveness								
1 (very ineffective)	5	3.11	1	−0.63	1	−0.40	1	−1.43
2	27	3.28	11	0.43	10	0.41	11	−0.26
3	69	3.34	24	0.08	23	−0.18	25	−0.25
4	76	3.49	32	0.09	31	−0.12	38	−0.34
5	91	3.64	37	0.34	35	0.13	40	0.04
6 (very effective)	67	3.81	36	0.51	34	0.25	41	0.41

Note: Value-added measures are standardized then averaged across math and reading.

better than *effective*. The remaining columns show a relatively large spread in value added between the lowest and highest categories of the two performance ratings.[15]

Question 3: What factors predict a teacher's official personnel evaluation score being higher or lower than predicted by the low-stakes rating?

Our final analysis considers whether teacher, principal, or school characteristics help predict gaps between a teacher's high-stakes rating and what rating would be predicted given the low-stakes ratings the principal assigned that teacher. For illustrative purposes, we proceed in two steps.[16] First, we regress the high-stakes rating factor scores (standardized) on the low-stakes rating factor scores (standardized). Points falling above this regression line (positive residuals) identify teachers whose high-stakes ratings are higher than would be predicted by their low-stakes ratings. Points falling below the line (negative residuals) indicate teachers whose high-stakes ratings are lower than would be predicted by their low-stakes ratings. A regression of the residuals $(Y - \hat{Y})$ on teacher, school, and principal characteristics—our second step—thus identifies variables associated with having a greater (or lower) than predicted high-stakes evaluation rating, given the rating provided in the low-stakes interview.

15. Appendix figure A.3 shows the probability of receiving a score of "highly effective" on each high-stakes standard by quintiles of the drift-adjusted value-added measures (averaging math and reading together). The figure shows upward-sloping lines, though not particularly steep ones. For example, for Standard 2 (Knowledge of Learners), 57 percent of teachers in quintile 1 (lowest) of value added received the highest rating, while 76 percent received the highest rating in quintile 5. For all but one standard, more than 40 percent of the teachers in the bottom value-added quintile received the highest rating.
16. A virtually identical analysis could be done in a single step in which high-stakes ratings were regressed on low-stakes ratings and the other characteristics.

Both steps use ordinary least squares.[17] The second stage includes a variety of teacher characteristics (sex, race/ethnicity, highest degree, experience level), school characteristics (fraction of students who are black, fraction who are free/reduced-price lunch eligible, enrollment size, school level, Florida accountability grade[18]), and principal characteristics (sex, race/ethnicity, highest degree, years in school) from administrative data. As measures of teacher performance, we also include teacher days absent in 2011–12 in some models and teacher value-added scores (applying empirical Bayes shrinkage and averaging over math and reading). Standard errors in the second stage are clustered at the school level.

Table 6 gives the second-stage results. Positive coefficients indicate that a teacher with a given characteristic scored better on the high-stakes evaluation than would have been predicted given the rating supplied by the principal in the low-stakes evaluation. Column 1 shows teacher characteristics only; columns 2 and 3 add school and principal characteristics. Columns 4 and 5 show results for teacher absences with and without other teacher, school, and principal variables. Columns 6 and 7 show results for the drift-adjusted value-added measure with and without the other variables.[19] We also estimated models using the average and 2011–12 value-added measures, which we do not show. In no cases were those measures significant predictors of the residual.

We examine teacher characteristics first, finding that black and Hispanic teachers receive lower high-stakes evaluations than would be predicted. The magnitude for both coefficients is approximately –0.3 standard deviation. Sex and highest degree are uncorrelated with the residual. Experience displays a kind of U-shape: novice teachers and teachers with twenty-one or more years of experience in the district both do better than would be predicted, as compared with teachers with intermediate levels of experience. The coefficient for novice teachers (i.e., those with fewer than two years of experience) is approximately twice as large as that for teachers with twenty-one or more years. This coefficient may indicate that principals inflate the scores of low-performing beginning teachers to encourage them or because these teachers enjoy the fewest job protections and are thus the most likely to be negatively affected by a low evaluation score at a time when they are still learning the job.[20]

In other untabulated analysis, we add school fixed effects to the model shown in column 1 of table 6, then test the null hypothesis that all of the fixed effects are equal to 0 in a joint *F*-test. We can reject this null hypothesis at the 0.01 level, which indicates that principals in some schools are more likely than others to have larger personnel evaluation scores than we would predict on the basis of the low-stakes ratings. Yet when we add school and principal characteristics in columns 2 and 3, we do not find much evidence that observable characteristics explain these idiosyncrasies. High school

17. Estimates from a feasible generalized least squares procedure to take into account the estimated nature of the dependent variable (Lewis and Linzer 2005) produced nearly identical results, so for simplicity we present the ordinary least squares results.
18. Because of sample size issues, we group D and F schools and B and C schools together. A is the omitted category.
19. Including teacher absences or not has no effect on the value-added coefficient.
20. In untabulated analyses in which we changed the categorization of the experience variable, we found that teachers with two years of experience (who also could not have tenure) did not have the same apparent inflation as teachers with less than two years, suggesting that the absence of job protections is not the sole reason that beginning teachers receive higher high-stakes evaluation scores than their low-stakes scores would predict.

Table 6. Predicting Residual of High-Stakes Score Regressed on Low-Stakes Interview Ratings

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Teacher Characteristics							
Female	0.041	0.125	0.133		0.151		0.103
	(0.104)	(0.106)	(0.110)		(0.109)		(0.151)
Black	−0.365***	−0.270*	−0.281**		−0.270*		−0.390*
	(0.129)	(0.138)	(0.139)		(0.140)		(0.209)
Hispanic	−0.272**	−0.237*	−0.242*		−0.244*		−0.092
	(0.131)	(0.130)	(0.133)		(0.132)		(0.180)
Has masters degree or higher	0.052	0.046	0.077		0.096		0.133
	(0.089)	(0.088)	(0.086)		(0.087)		(0.132)
Has 0–1 years experience in the district	0.776***	0.776*	0.580		0.583		–
	(0.280)	(0.475)	(0.360)		(0.382)		
Has 2–5 years experience in the district	−0.009	0.006	0.032		0.037		−0.128
	(0.148)	(0.145)	(0.138)		(0.138)		(0.228)
Has 6–9 years experience in the district	0.138	0.112	0.159		0.160		−0.018
	(0.139)	(0.134)	(0.141)		(0.141)		(0.174)
Has 21+ years experience in the district	0.319***	0.234**	0.262**		0.277**		0.040
	(0.108)	(0.108)	(0.115)		(0.115)		(0.164)
Teacher Performance							
Teacher's days absent				−0.005**	−0.008**		−0.003*
				(0.002)	(0.003)		(0.002)
Drift-adjusted value-added (math and reading averaged)						0.052	0.111*
						(0.052)	(0.063)
School Characteristics							
Fraction black students		−0.020	0.174		0.198		0.707*
		(0.239)	(0.249)		(0.249)		(0.357)
Fraction FRPL students		−0.475	−0.418		−0.505		0.181
		(0.564)	(0.570)		(0.570)		(0.773)
School enrollment (in 100s)		−0.007	−0.012		−0.011		0.024
		(0.010)	(0.010)		(0.010)		(0.023)
Middle school		0.182	0.108		0.127		0.118
		(0.152)	(0.153)		(0.152)		(0.173)
High school		0.396*	0.397*		0.385		−0.132
		(0.208)	(0.230)		(0.232)		(0.448)
D or F school accountability grade		0.138	0.193		0.183		0.021
		(0.213)	(0.224)		(0.225)		(0.282)
B or C school accountability grade		−0.039	0.036		0.025		−0.146
		(0.165)	(0.169)		(0.168)		(0.256)
Principal Characteristics							
Female			0.022		0.014		0.104
			(0.143)		(0.142)		(0.212)
Black principal			−0.455***		−0.449***		−0.621***
			(0.165)		(0.167)		(0.216)
Hispanic			−0.103		−0.084		0.158
			(0.136)		(0.134)		(0.230)
Has 2–3 years in this school			0.174		0.170		0.247
			(0.174)		(0.173)		(0.236)
Has 4–7 years in this school			0.192		0.204		0.058
			(0.185)		(0.185)		(0.256)
Has 8+ years in this school			−0.051		−0.058		0.091
			(0.291)		(0.294)		(0.518)
Has doctorate			0.002		−0.004		0.077
			(0.131)		(0.130)		(0.236)
Constant	0.014	0.200	0.118	0.019	0.162	−0.099	−0.747
	(0.146)	(0.439)	(0.488)	(0.066)	(0.498)	(0.086)	(0.676)
Observations	304	301	298	304	298	153	153
Adjusted R^2	0.059	0.081	0.100	0.005	0.116	0.000	0.079

Notes: Standard errors in parentheses, clustered by school.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

principals give more positive ratings, and black principals give lower ratings than would be predicted, but no other associations are statistically significant.

Columns 4 through 7 show that teacher absences are a consistent predictor of the residual irrespective of what other variables are included in the model. Principals give slightly lower-than-predicted ratings to teachers with larger numbers of work absences (every ten days is associated with a decrease of 0.03 to 0.08 standard deviation), consistent with the idea that principals place a premium on work effort in assigning teacher ratings on high-stakes assessments.[21] Value added is not associated with the residual value in the uncontrolled model but positively predicts the residual once other covariates are added ($\beta = 0.11$). That is, teachers with higher value added are assigned higher high-stakes ratings than predicted by their principals' low-stakes assessments of their performance.

Note that even in the model with the full set of covariates (column 7), we find that black teachers receive lower high-stakes evaluations than would be predicted (the coefficient for Hispanic teachers is no longer statistically significant), which may suggest a potential racial bias. In results not shown, we added school fixed effects to the model in column 7 (dropping school and principal characteristics) to further assess this possibility. The coefficient for black teachers remains negative but is somewhat attenuated ($\beta = -0.31$) and is no longer statistically significant at conventional levels ($p = 0.19$). Given the small number of teachers per school in these models, this analysis can be considered inconclusive but is suggestive that racial differences in teachers' evaluation ratings warrant attention in future research.[22]

5. DISCUSSION AND CONCLUSIONS

In an education policy era in which teacher evaluation is becoming central to school and district decision making, developing a deeper understanding of the capacity of school principals for facilitating high-quality teacher evaluations is critical. This analysis draws on unique data that pair high- and low-stakes evaluations of teachers by their principals with value-added measures of teachers' effectiveness. Findings reveal that principal ratings in both low- and high-stakes environments are negatively skewed and suggest that principals give lower ratings in low-stakes interviews with researchers than on high-stakes evaluation instruments. Principals rarely give low ratings on the high-stakes instrument, even when they report during low-stakes interviews that a teacher is ineffective on key performance dimensions.

Despite the fact that principals overwhelmingly rated teachers as effective or highly effective across standards on the high-stakes assessment, we find evidence that which of these categories they chose for a teacher revealed information about their true performance assessment. The correlation between the performance factors underlying the high- and low-stakes assessments correlated at 0.62. Moreover, while some in-class

21. We tested for nonlinearities in this relationship by including a squared absences term, but it was not statistically significant.

22. Given evidence that racial and ethnic congruence between teachers and principals influences their interactions (Grissom and Keiser 2011), we tested whether such congruence is associated with evaluation scores. Unexpectedly, we found that black teachers receive statistically lower-than-predicted scores in schools with black principals than in schools with non-black principals. A deeper look is beyond the scope of the present study but points in an interesting direction for future work.

ratings of teacher skills from the low-stakes interviews tended to be somewhat more predictive of value added, the two summative factor ratings correlated similarly well with teacher value added in both subjects. In other words, despite leaving ratings of 1 or 2 virtually unused, teachers' ratings on the high-stakes IPEGS instrument correlated as well with measures of their impacts on student achievement as the more dispersed measures provided in the low-stakes interviews.

We also find that there are patterns in principals' propensities to inflate (or deflate) ratings on the high-stakes instrument that suggest that in the high-stakes setting principals respond to pressures and do not give unfiltered evaluations of teachers. In particular, they appear to give higher-than-predicted ratings to beginning teachers, though we should caution that the passage of Florida's Student Success Act (Senate Bill 736) in 2011, which abolished the possibility of tenure protections for newly hired teachers, may mean that this finding is a "cohort effect"—these teachers were the first impacted by the new law—rather than a novice teacher effect that we might observe more generally. Results also suggest that there are some differences in principals' inflation/deflation probabilities across schools, but with available data, we are generally unable to identify the factors contributing to these differences.

Taken together, our results point to some important considerations for designers of teacher evaluation systems. Prior studies finding that principals in low-stakes settings generally could identify which teachers in their schools are more able to raise student test scores have been interpreted as evidence that principals have capacity to differentiate their higher- and lower-performing teachers (Jacob and Lefgren 2008; Harris and Sass 2014). Our results suggest that principals' ratings in high-stakes environments reflect such differentiation as well, to a degree similar to low-stakes ratings. Still, principals face apparently strong pressures to skew their ratings of teachers away from their true beliefs about that performance when there are stakes attached. This inflation of ratings, at least in this district, appears on one hand to be structural in the sense that principals simply tend not to give low ratings on any standards, on average. On the other hand, principals are also more likely to inflate with some kinds of teachers (e.g., new ones) than others. Presumably, getting principals to give "truer" ratings that also make greater use of lower rating categories would facilitate more accurate feedback to teachers, provide greater incentives for improvement for low performers, and make it more likely that struggling teachers who do not improve exit the system (Sartain, Stoelinga, and Brown. 2011; Drake et al. 2016; Sartain and Steinberg, 2016). Meeting these goals may require changes to evaluation processes or principal professional development related to evaluation processes. Requiring frequent observations or rigorous evidence gathering, using detailed observation rubrics that clearly describe performance expectations, and training and coaching principals both to conduct high-quality evaluations consistent with district goals and have constructive feedback conversations, are strategies for improving the quality-of-performance assessment data and making it useful for district and school decision making (Grissom et al. 2017).

These conclusions, however, are tempered by a number of limitations. The number of teachers for whom we are able to compare high- and low-stakes ratings is small. The setting for the study is one large, urban district, and we do not know the degree to which these results generalize to other types of districts or to districts utilizing other modes of teacher evaluation. Also, the 2011–12 school year was the first year that Florida law

required high-stakes evaluation for all teachers in the state. Principals may have changed how they evaluated teachers in subsequent years as they learned to implement the new evaluation system.[23] Perhaps more importantly, there is misalignment between the constructs assessed on the two instruments, given the differences in items used. Another potentially important difference between the two instruments is the difference in the number of scale points principals could utilize; the low-stakes assessment's six-point scale may have helped principals feel more comfortable utilizing lower scores than the four-point scale used on the high-stakes assessment. We cannot be certain how differences in constructs and rating scales between the two kinds of teacher assessments may have impacted the results.

This study suggests that researchers should pay greater attention to the cognitive and relational processes that surround principals' subjective evaluations of teachers. Future work should also delve into principals' evaluations of teachers in other contexts. Of particular interest is the predictive validity of high-stakes evaluation ratings in systems that invest heavily in frequent observations by multiple raters (see Goldring et al. 2015), which may show a different pattern of results. Still, our results contribute to our growing understanding of the roles of school leaders in teacher evaluation, which will only become more important as teacher evaluation data increasingly drive instructional and personnel decisions in schools (Neumerski et al. 2014). In particular, this study has shown that even though high-stakes ratings are severely compressed, they include information both about teachers' success at raising test scores and about teachers' other valued contributions that are measured, likely more accurately, by low-stakes assessments.

ACKNOWLEDGMENTS

The authors thank Julie Cohen, Mimi Engel, the anonymous referees, and attendees of the 2013 Association for Public Policy Analysis and Management and 2014 Association for Education Finance and Policy annual meetings for helpful comments on earlier drafts. This research was supported by a grant from the Institute of Education Sciences at the U.S. Department of Education (R305A100286).

REFERENCES

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review* 104(9):2593–2632. doi:10.1257/aer.104.9.2593.

Darling-Hammond, Linda, Audrey Amrein-Beardsley, Edward Haertel, and Jesse Rothstein. 2012. Evaluating teacher evaluation. *Phi Delta Kappan* 93(6):8–15. doi:10.1177/003172171209300603.

Doherty, Kathryn M., and Sandi Jacobs. 2013. *Connect the dots: Using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality.

Drake, Timothy A., Ellen Goldring, Jason A. Grissom, Marisa Cannata, Christine M. Neumerski, Mollie Rubin, and Patrick Schuermann. 2016. Development or dismissal? Exploring principals'

23. Some evidence suggests that principals statewide differentiated evaluations just slightly more in the second year than in the first year (Sawchuk 2013).

use of teacher effectiveness data. In *Improving teacher evaluation systems: Making the most of multiple measures*, edited by Jason A. Grissom and Peter Youngs, pp. 116–130. New York: Teachers College Press.

Epstein, Joyce L. 1985. A question of merit: Principals' and parents' evaluations of teachers. *Educational Researcher* 14(7):3–10. doi:10.3102/0013189X014007003.

Frase, Larry E., and William Streshly. 1994. Lack of accuracy, feedback, and commitment in teacher evaluation. *Journal of Personnel Evaluation in Education* 8(1):47–57. doi:10.1007/BF00972709.

Glazerman, Steven, Susanna Loeb, Dan Goldhaber, Douglas Staiger, Stephen Raudenbush, and Grover Whitehurst. 2010. *Evaluating teachers: The important role of value-added*. Washington, DC: Brookings Institution.

Grissom, Jason A., Demetra Kalogrides, and Susanna Loeb. 2015. Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis* 37(1):3–28. doi:10.3102/0162373714523831.

Grissom, Jason A., and Lael Keiser. 2011. A supervisor like me: Race, representation, and the satisfaction and turnover decisions of public sector employees. *Journal of Policy Analysis and Management* 30(3):557–580. doi:10.1002/pam.20579.

Grissom, Jason A., and Susanna Loeb. 2011. Triangulating principal effectiveness: How perspectives of parents, teachers, and assistant principals identify the central importance of managerial skills. *American Educational Research Journal* 48(5):1091–1123. doi:10.3102/0002831211402663.

Grissom, Jason A., Susanna Loeb, and Benjamin Master. 2013. Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher* 42(8):433–444. doi:10.3102/0013189X13510020.

Grissom, Jason A., Mollie Rubin, Christine Neumerski, Marisa Cannata, Timothy Drake, Ellen Goldring, and Patrick Schuermann. 2017. Central office supports for data-driven talent management decisions: Evidence from the implementation of new systems for measuring teacher effectiveness. *Educational Researcher* 46(1):21–32. doi:10.3102/0013189X17694164.

Gulliksen, Harold. 1987. *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum.

Harris, Douglas N., W. Kyle Ingle, and Stacey A. Rutledge. 2014. How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal* 51(1):73–112. doi:10.3102/0002831213517130.

Harris, Douglas N., and Tim R. Sass. 2014. Skills, productivity and the evaluation of teacher performance. *Economics of Education Review* 40:183–204. doi:10.1016/j.econedurev.2014.03.002.

Jacob, Brian A., and Lars Lefgren. 2005. Principals as agents: Subjective performance measurement in education. NBER Working Paper No. 11463.

Jacob, Brian A., and Lars Lefgren. 2008. Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics* 26(1):101–136. doi:10.1086/522974.

Kimball, Steven M., Brad White, Anthony T. Milanowski, and Geoffrey Borman. 2004. Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education* 79(4):54–78. doi:10.1207/s15327930pje7904_4.

Lewis, Jeffrey B., and Drew A. Linzer. 2005. Estimating regression models in which the dependent variable is based on estimates. *Political Analysis* 13(4):345–364.

MacLeod, W. Bentley. 2003. Optimal contracting with subjective evaluation. *American Economic Review* 93(1):216–240. doi:10.1257/000282803321455232.

Medley, Donald M., and Homer Coker. 1987. The accuracy of principals' judgments of teacher performance. *Journal of Educational Research* 80(4):242–247. doi:10.1080/00220671.1987.10885759.

Neumerski, Christine, Jason A. Grissom, Ellen Goldring, Marisa Cannata, Timothy Drake, Mollie Rubin, and Patrick Schuermann. 2014. Inside teacher evaluation systems: Shifting the role of the principal as instructional leader. Paper presented at the Thirty-ninth Annual Conference of the Association for Education Finance and Policy, San Antonio, TX, March.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2012. Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review* 102(7):3184–3213. doi:10.1257/aer.102.7.3184.

Rothstein, Jesse. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics* 125(1):175–214. doi:10.1162/qjec.2010.125.1.175.

Sartain, Lauren, Sara Ray Stoelinga, and Eric R. Brown. 2011. *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Chicago: Consortium on Chicago School Research.

Sartain, Lauren, and Matthew Steinberg. 2016. Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago Public Schools. *Journal of Human Resources* 51(3):615–655.

Sawchuk, Stephen. 2013. *Teachers' ratings still high despite new measures*. Available www.edweek.org/ew/articles/2013/02/06/20evaluate_ep.h32.html. Accessed 18 August 2016.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, David Keeling, Joan Schunck, Ann Palcisco, and Kelli Morgan. 2009. *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.

APPENDIX A: DETAILS ON BAYESIAN SHRINKAGE

Our estimated teacher effect ($\hat{\delta}_t$) is the sum of a "true" teacher effect (δ_t) plus some measurement error[24]:

$$\hat{\delta}_t = \delta_t + \varepsilon_t. \tag{A.1}$$

The empirical Bayes estimate of a teacher's effect is a weighted average of his estimated fixed effect and the average fixed effect in the population where the weight, λ_t, is a function of the precision of each teacher's fixed effect and therefore varies by t. The less precise the estimate, the more we weight the mean. The more precise the estimate, the more we weight the estimate and the less we weight the mean. Similarly, the more variable the true score (holding the precision of the estimate constant) the less we weight the mean, and the less variable the true score, the more we weight the mean

24. Here we make the classical errors in variables assumption, assuming that measurement error is not associated with an unobserved explanatory variable.

Table A.1. Correlations among Value-Added Measures

	(1)	(2)	(3)	(4)	(5)	(6)
(1) Average value added, all years (math)	1					
(2) Average value added, all years (reading)	0.70	1				
(3) Value added, 2011–12 (math)	0.62	0.55	1			
(4) Value added, 2011–12 (reading)	0.49	0.57	0.58	1		
(5) Drift-adjusted value added, 2011–12 (math)	0.62	0.57	0.46	0.34	1	
(6) Drift-adjusted value added, 2011–12 (reading)	0.67	0.41	0.51	0.24	0.86	1

Table A.2. Spearman Correlations within Ratings Variables

Panel A: High-Stakes Evaluation Instrument

Standard		2	3	4	5	6	7	8
2	Knowledge of Learners	1						
3	Instructional Planning	0.49	1					
4	Instructional Delivery and Engagement	0.53	0.50	1				
5	Assessment	0.44	0.47	0.45	1			
6	Communication	0.39	0.36	0.35	0.37	1		
7	Professionalism	0.35	0.36	0.36	0.38	0.45	1	
8	Learning Environment	0.52	0.47	0.55	0.40	0.40	0.37	1

Panel B: Low-Stakes Interview Instrument

Items		1	2	3	4	5	6	7	8	9	10
	"In-Class" Items										
1	Getting High Test Performance	1									
2	Improving Critical Thinking	0.83	1								
3	Motivation	0.67	0.76	1							
4	Building Interpersonal Skills	0.56	0.62	0.80	1						
5	Overall In-Class Effectiveness	0.80	0.85	0.84	0.75	1					
	"Out-of-Class" Items										
6	Building Staff Relationships	0.47	0.50	0.58	0.64	0.62	1				
7	Supporting Colleague Instruction	0.59	0.64	0.67	0.62	0.73	0.78	1			
8	Helping with School Leadership/Management	0.50	0.55	0.57	0.58	0.63	0.74	0.80	1		
9	Building Community Relationships	0.47	0.51	0.58	0.64	0.60	0.71	0.70	0.76	1	
10	Overall "Out-of-Class" Effectiveness	0.59	0.62	0.65	0.66	0.72	0.81	0.84	0.87	0.84	1

(assuming the true score is probably close to the mean). The weight, λ_j, should give the proportion of the variance in what we observe that is due to the variance in the true score relative to the variance due to both the variance in the true score and precision of the estimate. This more efficient estimator of teacher quality is generated by:

$$E(\delta_t|\hat{\delta}_t) = (1-\lambda_t)\left(\bar{\delta}\right) + (\lambda_t) * \hat{\delta}_t, \tag{A.2}$$

$$\text{where } \lambda_t = \frac{(\sigma_\delta)^2}{(\sigma_{\varepsilon t})^2 + (\sigma_\delta)^2}. \tag{A.3}$$

Thus, the term λ_t can be interpreted as the proportion of total variation in the teacher effects that is attributable to true differences between teachers. The terms in equation

Table A.3. Spearman Correlations among Low- and High-Stakes Rating Items

	High-Stakes Evaluation Ratings							
Low-Stakes Ratings	**Standard 2**	**Standard 3**	**Standard 4**	**Standard 5**	**Standard 6**	**Standard 7**	**Standard 8**	**Mean**
Getting High Test Performance	0.36	0.38	0.43	0.40	0.16	0.24	0.36	0.46
Improving Critical Thinking	0.44	0.38	0.48	0.44	0.17	0.28	0.39	0.51
Motivation	0.41	0.30	0.41	0.38	0.21	0.29	0.38	0.47
Building Interpersonal Skills	0.36	0.24	0.31	0.22	0.20	0.24	0.32	0.38
Overall In-Class Effectiveness	0.47	0.39	0.50	0.42	0.23	0.31	0.44	0.55
Building Staff Relationships	0.31	0.28	0.31	0.27	0.21	0.34	0.29	0.41
Supporting Colleague Instruction	0.37	0.33	0.40	0.32	0.22	0.38	0.35	0.47
Helping with School Leadership/Management	0.30	0.30	0.30	0.30	0.24	0.36	0.31	0.42
Building Community Relationships	0.34	0.28	0.27	0.26	0.25	0.32	0.29	0.40
Overall Out-of-Class Effectiveness	0.39	0.37	0.39	0.32	0.26	0.39	0.38	0.50

Note: Labels for Standards 2 through 8 are omitted, but they correspond to the standards numbered in table 1.

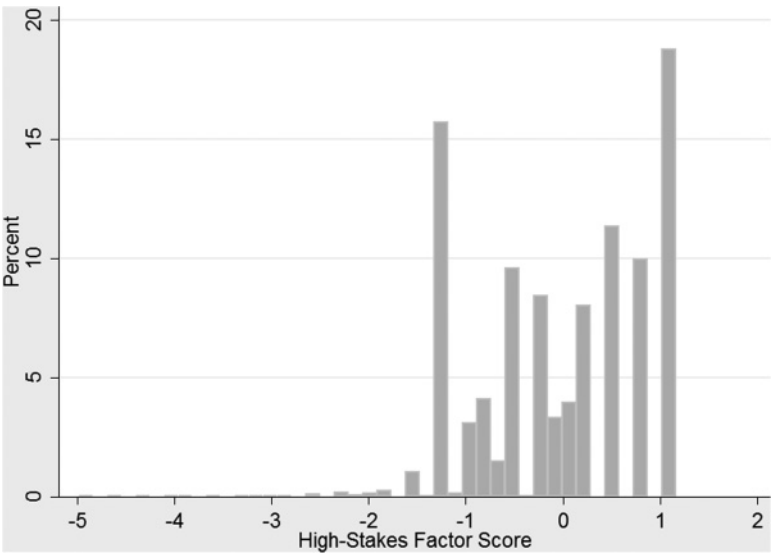


Figure A.1. Distribution of High-Stakes Evaluation Factor Scores.

A.3 are unknown so are estimated with sample analogs,

$$(\hat{\sigma}_{\varepsilon t})^2 = var\left(\hat{\delta}_{\varepsilon t}\right). \tag{A.4}$$

The analog is the square of the standard error of the teacher fixed effects. The variance of the true fixed effect is determined by:

$$(\sigma_\delta)^2 = (\hat{\sigma}_\delta)^2 - mean\,(\hat{\sigma}_\varepsilon)^2, \tag{A.5}$$

where $(\hat{\sigma}_\delta)^2$ is the variance of the estimated teacher fixed effects (Jacob and Lefgren 2005; Grissom, Kalogrides, and Loeb 2015).

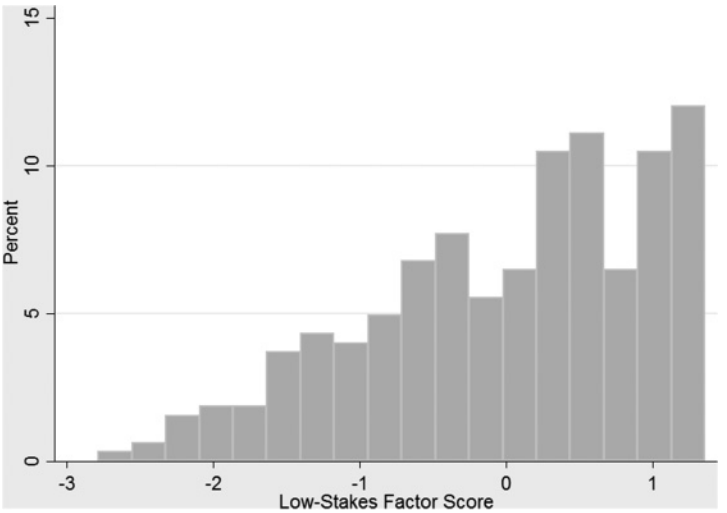
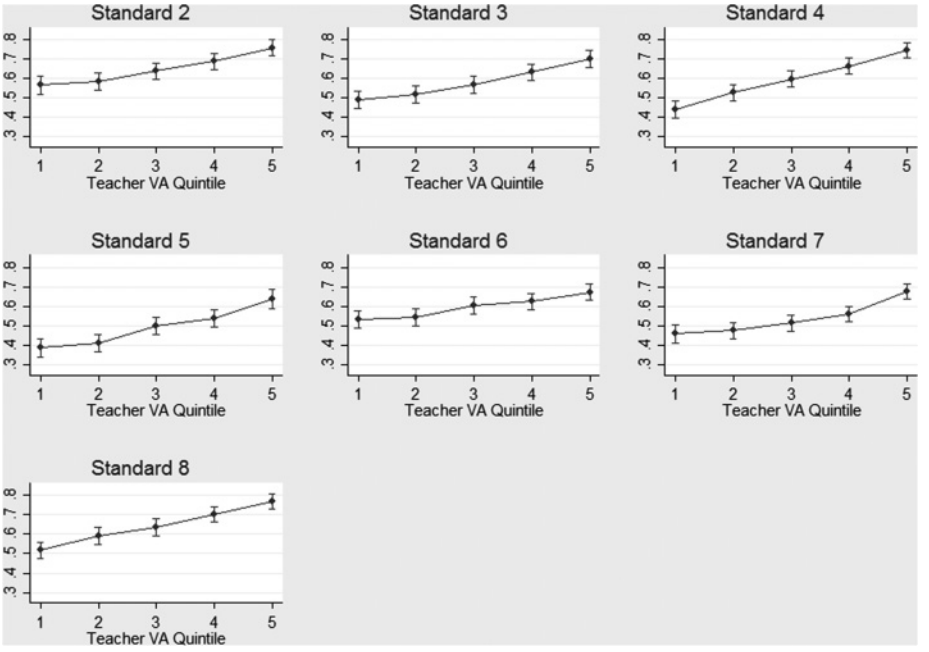


Figure A.2. Distribution of Low-Stakes Interview Ratings Factor Scores.



Note: Drift-adjusted value-added measures shown.

Figure A.3. Probability of Being Scored as Highly Effective by Value-Added Quintile, by High-Stakes Standard.