Staffing for Success: Linking Teacher Evaluation and School Personnel Management in Practice

Benjamin Master (bmaster@stanford.edu)

Stanford University

Abstract

Teacher evaluation is at the center of current education policy reform. Most evaluation systems rely at least in part on principals' assessments of teachers, and their discretionary judgments carry substantial weight. However, we know relatively little about what they value when determining evaluations and high stakes personnel decisions. I leverage unique data from a public charter school district to explore the extent to which school administrators' formative evaluations of teachers align with teacher and school effectiveness and predict future personnel decisions. While previous research has examined administrators' subjective evaluations of teachers in surveys and in practice, this study links a detailed evaluation instrument in practice with different types of personnel decisions to provide new insights into administrator decision-making. The results indicate that formative mid-year ratings – shared by administrators with teachers – are strongly associated with end-of-year dismissal and promotion decisions. I use an exploratory factor analysis to identify four distinct components of administrators' feedback to teachers and show that different components predict different types of personnel decisions in schools. I also find that different teacher evaluation factors are associated with individual versus school-wide student achievement gains. The results suggest the importance of accounting for multiple aspects of teacher performance in evaluation systems that are meant to inform multiple types of personnel decisions.

I. Introduction

The Federal Race to the Top Initiative has spurred development and implementation of new teacher evaluation systems as a key lever for improving school effectiveness and raising student achievement. Evaluation systems may improve the quality of teaching via two key mechanisms. First, they may identify and promote effective teaching practices that help teachers to improve (Taylor and Tyler, 2011). Second, they may facilitate personnel practices and policies that support the retention of more effective teachers and the dismissal of less effective teachers, as well as more optimal assignment of teachers to jobs in which they can have the most positive effect (Boyd et. al, 2010; Goldhaber and Theobald, 2011; Rockoff et. al, 2011).

To accomplish either of these aims, educators must leverage measures of teacher effectiveness without inadvertently neglecting important contributions that occur outside the scope of measurement. That is, evaluation systems will not be as effective if the evaluation measures used miss important components of teaching that could aid in teacher improvement or more effective personnel practices. Thus far, most emphasis in current reforms has been devoted to value-added measures of teacher effectiveness based on student test performance. While these measures address central aspects of teachers' work, they may do a poor job of accounting for teacher impacts on valued student outcomes other than annual tested achievement, such as motivation, character development, or achievement outside the scope of standardized tests. They may also miss valuable teacher contributions that occur outside of regular classroom practice, such as organizational leadership, relationships with students and families, or collaboration with peers.

Observational protocols of teaching practices, including the Framework for Effective Teaching (FFT), the Classroom Assessment Scoring System (CLASS), and the Protocol for Language Arts Teaching Observations (PLATO), have also received substantial attention in new evaluation systems. For instance, local variations of the FFT have been leveraged by evaluation systems in

Denver, Cincinnati and New Haven public schools. Researchers studying teacher evaluation have primarily focused on developing these protocols based on conceptions of good teaching and assessed the relationship between the measures from these evaluation tools and value added measures of teachers' impact on student achievement growth (Gates Foundation, 2012; Kane et. al, 2011; Grossman et. al, 2010; Hamre and Pianta, 2005).

Prior to the recent wave of new measures, and even today, most teacher evaluation is based on principal assessment of teachers. These evaluations are far less specific than either the value added or observational protocol measures but this lack of specificity may allow for a fuller view of teaching. Both value-added measures and direct observational protocols may fail to measure important aspects of teacher performance. If this oversight is the case, teachers may receive the wrong signals about how best to improve their own performance. Similarly, personnel policies that are determined by these measures may misapply high stakes consequences such as teacher promotions, role assignments and dismissals.

One way of exploring the diversity of teacher contributions and better understanding the extent to which value-added and observational protocols are capturing quality teaching is to investigate the practices of local school administrators who engage in subjective or standards-based teacher evaluation. Subjective teacher evaluation refers to holistic administrator judgments based on flexible criteria. Standard-based teacher evaluations link those judgments to a more fixed set of standards that define a competency model of effective teaching (Heneman et. al, 2006). Many emerging teacher evaluation systems leverage administrator perspectives of one or both types, at least in part (e.g. Denver's ProComp, New Haven Public Schools' TEVAL). Administrators' more holistic judgments of teachers are valued in part because they can capture aspects of job performance that may be missed by formal evaluation instruments. Administrators are also essential actors because they play a critical role as mentors in formative evaluation systems.

Research on administrators' evaluations of teachers has generally focused on whether they predict value added measures of teacher effectiveness. For instance, Jacob and Lefgren (2008) use survey measures to find that principals can directly identify very high and low value added teachers. A substantial body of research has also linked subjective and standards-based principal evaluations with objective teacher value added measures in practice (e.g. Holtzapple, 2003; Gallagher, 2004; Kimball et al., 2004; Milanowski, 2004; Rockoff and Speroni, 2011). As a result, we know that principals are capable of partially identifying teacher value added performance, but we do not know much about how well principals' priorities are reflected by value added measures.

A few studies explore additional teacher characteristics that may be valued by administrators. Harris and Sass (2009) survey principals in order to identify a variety of teacher traits that they believe are important to teaching. Among these traits, they find that principals' assessments of teachers' subject knowledge, teaching skill, and intelligence are associated with value added effectiveness, while their assessments of teachers' interpersonal skills are not. In a related vein, Jacob and Lefgren (2008) find that survey-reported principal ratings of teachers are substantially better predictors of parent requests for teachers than value added measures. Jacob and Walsh (2010) identify associations between subjective evaluations determined in practice by administrators and observable teacher characteristics, including attendance, experience, and credentials. They also observe relationships between subjective ratings and some teacher assignments. These studies do not, however, examine the import of specific evaluation criteria in contexts where local administrators have clear autonomy over personnel decisions. This is important because we may learn more about the relative priorities administrators ascribe to different evaluation criteria when they are fully responsible for making tradeoffs between them.

In current practice, administrators' subjective evaluations do a poor job of distinguishing between effective and ineffective teachers (Weisburg et al., 2009). However, this may be a product

of the nature of existing evaluations systems. Rules surrounding administrators' capacity to conduct evaluations or implement high stakes differentiation are often complex, ambiguous, and limiting — either overall or in particular aspects of evaluation (Hess and Loup, 2008; Price, 2009; Ballou, 2000). Under these circumstances, they may have little incentive to accurately assess teacher performance.

Administrators do take action to dismiss less effective teachers and promote more effective teachers when they are empowered to do so (Jacob 2010; Chingos and West, 2011; Rockoff et. al, 2011). Thus, it seems likely that subjective evaluations by administrators who have greater autonomy will yield more credible insights into what they actually value. Moreover, because emerging evaluation reforms are providing schools greater discretion in teacher personnel management, it is increasingly important to understand administrators' perspectives in this regard.

A better understanding of the teacher contributions that administrators consider in their personnel decisions may improve the design of emerging evaluation systems. Investigations of administrator practices can help to illuminate not only what they care about, but what they can observe and how they act upon those observations. While the measures utilized in teacher evaluation system are ultimately discretionary, additional insights into administrators' perspectives and professional judgments offer two key benefits. First, they can inform the selection of measures and professional standards considered in evaluations. Second, they may identify common disconnects between desirable standards and the priorities of local school leaders who will be responsible for their implementation.

In this study I leverage unique data from a public charter school district to explore the extent to which school administrators' formative evaluations of teachers align with teacher and school effectiveness and predict future personnel decisions. While previous research has examined administrators' subjective evaluations of teachers in surveys and in practice, this study links a detailed evaluation instrument in practice with different types of personnel decisions to provide new

insights into administrator decision-making. The results indicate that formative mid-year ratings – shared by administrators with teachers – are strongly associated with end-of-year dismissal and promotion decisions. I use an exploratory factor analysis to identify four distinct factors of administrators' feedback to teachers and show that different factors predict different types of personnel decisions in schools. I also find that different teacher evaluation factors are associated with individual versus school-wide student achievement gains. The results suggest the importance of accounting for multiple aspects of teacher performance in evaluation systems that are meant to inform multiple types of personnel decisions.

The remainder of the paper proceeds as follows. Section II summarizes my hypotheses and research questions. Section III provides relevant background information about the district of study, and details my data. Section IV specifies my empirical approach, while Section V presents the results. Section VI concludes with a discussion of implications and potential limitations.

II. Hypotheses and Research Questions

This study explores a variety of school personnel decisions, including teacher dismissals, voluntary teacher resignations, administrators' identification of likely candidates for future promotions, and administrators' actual promotion of teachers to school leadership roles. I link each of these to mid-year evaluative feedback provided by administrators to teachers. I also link teacher and school value added measures to the evaluations teachers receive. My goals are twofold. First, to examine the extent to which the formative feedback administrators provide to teachers is aligned with their actual personnel decisions and with future student achievement outcomes. Second, I aim to better understand the considerations that may be influencing their personnel decisions.

Hypotheses

In the following, I elaborate on each type of personnel decision in sequence, detailing the considerations that may be relevant to administrators. I also consider how administrators' evaluations of teachers may account for teacher and school value added effectiveness.

Dismissals. Research on teacher dismissals suggests that perceived value added performance is a factor weighed by administrators (Rockoff et al., 2011). Teacher attendance, experience, and credentials can also predict the likelihood of dismissals (Jacob and Walsh, 2010). Teacher dismissal decisions likely reflect administrators' observation that a teacher is unable to meet the minimum standard for performance in the school, and is failing students as a result. Skills that are considered essential to effective teaching may be lacking, or the teacher may be a poor fit for certain "non-negotiable" aspects of the school culture. Administrators will likely observe perceived teacher deficiencies in classroom practice and in teachers' instructional planning. In mid-year evaluations, administrators would probably communicate their concerns to these teachers in a desire to improve their performance, and might focus their feedback on foundational aspects of teaching.

Resignations. Voluntary teacher resignations likely reflect a wide variety of different motivations and circumstances that are beyond administrators' purview. However, administrators can exert substantial influence on teachers' voluntary resignation decisions (Rockoff et al., 2011; Balu et al., 2010). Thus, the population of resigning teachers will likely include more staff that administrators view as below average performers than the other way around. As a group, they will likely be rated worse, potentially along dimensions that are not so critical as to merit an outright dismissal.

School leadership promotions. Research on school leadership promotions suggests that administrators' decisions reflect teachers' value added effectiveness (Chingos and West, 2011), but it is unclear what else they value in practice. Decisions to promote teachers to Assistant Principal and Principal roles likely reflect administrators' observation that a teacher is demonstrably "on board" with school-wide systems, norms and values, and is capable of championing them. Assistant

Principals' responsibilities can vary substantially within schools, and in this district are bifurcated between "academic leadership" and "school culture" designations. Based in part on district-provided role descriptions, I expect that academic leadership promotions are based on teachers' ability to manage instructional practice at the school and to improve student achievement outcomes. School culture leadership promotions appear to reflect a more specialized skill set for managing student-facing norms and behavior, and relationships with students' parents. Promotions to both roles will likely be based on observations of teacher contributions both in and out of the classroom. Leadership skills and interpersonal relationships are likely to be a factor in both types of promotion.

Consideration for promotions. In addition to enacted promotion decisions, I examine administrators' internal evaluation of teachers' potential for future promotions. In general, I expect that near-term promotion candidates will possess the same mix of positive qualities as those who are actually promoted, but to a lesser degree. Candidates for potential promotion in the longer term (i.e. 3 to 5 years), however, are more likely relative newcomers to the school whose foundational skills and "buy in" to school norms are seen to reflect a potential for rapid improvement.

Expert Teachers. Teachers who are flagged for potential promotion to an "Expert Teacher" role are likely viewed as instructional experts. They may be deemed to be effective with students and seen as role models of instructional practice for other teachers. As such, their expertise should be demonstrated in their classroom practice. Expert Teacher designation likely reflects administrators' performance aspirations for the majority of regular classroom teachers in their school.

Value added effectiveness. Research suggests that teachers' individual effectiveness at raising student achievement is of interest to administrators and is something that they can identify, to an extent (Jacob and Lefgren, 2008). They leverage this information when it is made available (Rockoff et. al, 2011) and may weigh their determination of it heavily in their overall subjective appraisals of staff (Harris and Sass, 2009). Given this, administrators' perspective on teachers' effectiveness at raising

student achievement may be partially reflected in the formative evaluations they share with teachers. In addition, administrators may also be concerned with teachers' collaborative contributions towards school-wide effectiveness at raising student achievement, and may account for that in their formative evaluations as well.

However, since neither individual nor school-wide value added effectiveness are explicitly addressed in the formative evaluation instrument, any relationship between ratings and teacher or school-wide value added effectiveness will be indirect. Instead, as part of the evaluation instrument, a single evaluation indicator assessed the absolute level of observed student achievement by mid-year on quizzes and tests, relative to specific, ambitious performance targets. Administrators' evaluations of teachers in a particular school year may also be biased by their impressions of that teacher from prior school years, and this could either strengthen or weaken the association between their evaluations and value added in the same year. For all these reasons, it is possible that the formative ratings administrators provide will fail to accurately assess teachers' same-year contributions to either classroom or school-wide student achievement gains.

Research Questions

In order to test these hypotheses, I specifically address the following questions of interest:

1. Do overall ratings on formative mid-year teacher evaluations predict subsequent dismissal and promotion decisions by administrators?
2. Are there coherent and distinct factors within evaluative ratings that reflect different aspects of teacher performance?
3. Are different types of personnel decisions or anticipated personnel decisions predicted by different factors from the evaluative ratings?
4. Do either overall ratings or specific factors from the evaluations predict teacher value added performance in the same school year?

5. Do school aggregates of all teachers' evaluative ratings predict school-wide average value added performance in the same school year?

III. District Background and Data

My data come from a network of highly effective[1] public charter schools that operate alongside state public schools under a single centralized district management team (hereafter referred to as "the district").[2] District schools serve an over-subscribed, lottery-selected population of predominantly poor and minority students across grades K-12. 75% of district students qualify for free or reduced price lunch, and the student population is made up of 80% African American students and 19% Hispanic students.

District practices are unique in some important ways. The district's personnel management practices regarding teacher dismissals and promotions represent the autonomous decisions of local school principals, assistant principals, and district administrators – unrestricted by any external contracts or policies. Also of note, promotion to school leadership in this district includes an explicit expectation to continue some classroom teaching, while also actively leading and mentoring other staff. Finally, during the period of this study, the district was exploring the creation of a unique "Expert Teacher" promotion for skilled instructors who do not aspire to school leadership.

District data – including teacher characteristics, personnel decisions, and teachers' evaluative ratings – are available over three years, from school year (SY) 2008-2009 through SY 2010-2011. During this period, individual district schools participated for one or more years in a common, formative, teacher evaluation system. This system was implemented independently at each participating school but was based on a shared set of professional and teaching standards. District personnel decisions were not explicitly linked to the evaluation system. Details about each category of data provided by the district are provided in the following sub-sections.

Mid-Year Formative Evaluations

In January or February of each school year, administrators in participating schools documented formal mid-year teacher evaluations for the majority of their full-time teachers. These evaluations reflect principal and assistant principal judgments about teachers using a loosely defined rubric that covers 47 different indicators of performance so far that year. Each indicator is rated on either a 5-or-4-point Likert scale with ratings ranging from "Role Model" to "Needs Development".

Table 2 includes a complete list of the individual evaluation indicators used by the district, paraphrased for brevity.[3] It also shows the conceptual structure district actors used as a framework for evaluation. Each indicator is grouped into one of 7 dimensions of professional excellence: Achievement, Character, Instruction, Classroom Culture, Systems and Planning, Student and Family Relationships, and Personal Effectiveness. The evaluation standards used by the district reflect the internally negotiated instructional and organizational priorities of district and school leaders, and were reinforced during each school year as part of school leader and teacher professional development activities. However, no systematic norming or calibration efforts were made to ensure identical rating practices across schools.

School leaders rated their staff autonomously, based on their holistic impressions from the first half of the school year. Evaluation ratings were shared with each teacher as part of a formal mid-year review meeting, and teachers also completed a separate self-evaluation of the same criteria in preparation for that discussion. This activity was primarily intended to be formative in nature, and was not systematically tied to any specific high stakes personnel decisions.

As detailed in Table 1, a total of 747 teacher ratings were documented over the three year period of the study. 77% of full time teachers in participating schools received a fully documented mid-year review. 56% of these teacher ratings had either one (29%) or more than one (27%) missing score across the 47 indicators. District records indicate that this primarily occurred when evaluators felt they did not have enough information to determine a rating on a particular dimension. In order to

avoid dropping the entire evaluative record for these teachers, I employ an Iterative Chained Equations (ICE) approach to impute the missing indicators, using data from non-missing indicators and teacher demographic characteristics to arrive at predicted values.

Staff Demographic Data

Available staff demographic data linked to teachers and school years were provided by the district. These include teachers' gender, age, race, and the length of their tenure at the district. The average within-district tenure for teachers was 2.3 years. For most teachers, data were also available on their total years of teaching experience, including years outside of the district. However, because lifetime teaching experience data was missing in substantially non-random ways for 142 teachers, and in patterns that were correlated with some outcomes of interest, I do not use this data in the study. Nevertheless, it is notable that teachers who were in their first year of teaching in the district possessed an average prior teaching experience of 2.4 years. This reflects district policies of primarily recruiting teachers with at least some prior teaching experience, and is an important consideration when interpreting associations with local teacher experience.

High Stakes Personnel Decisions and Anticipated Personnel Decisions

For each year of the study, the district provided data on which teachers were either dismissed or promoted before the start of the following school year. Promotions were documented for two types of school leadership positions, related to academic and school culture leadership, as previously discussed. Finally, teachers' voluntary resignations were also documented for each year of the study.

In a separate category, the district also conducted a one-time internal census of local administrators' perspectives in SY 2009-2010. This census was meant to identify the pipeline of teachers who were plausible candidates for promotion in future school years. Specifically, school leaders identified teachers that they thought had the potential to be effective school leaders within the next 2 years, teachers that they thought might be effective school leaders within the next 3 to 5

years, and candidate teachers for promotion to a new "Expert Teacher" role. The Expert Teacher role had not yet been instituted by the district, but was a hypothesized new role meant to recognize and reward teachers with strong instructional and coaching skills who were nevertheless not ideal candidates for school leadership. This internal census of administrators' perspectives was non-binding, and was not discussed at teachers' formative evaluation meetings.

Table 1 details the frequency of each type of anticipated and actual personnel decision. Internal teacher promotion rates varied substantially from year to year, as a function of leadership turnover and external hiring, while teacher dismissal and resignation rates were more consistent over time.

Teacher Value Added Estimates

The district began developing the capacity to generate value added measures of student achievement gains beginning in SY 2009-2010. For a variety of reasons, they opted to compare their students' performance on state standardized tests relative to achievement gains of students in the same grade level who were from large and geographically close external benchmark districts. They instituted data sharing agreements with those districts to facilitate this. They then contracted with an external vendor to develop their value added model.

The district did not begin systematically linking individual students to their teachers to facilitate the creation of individual teacher value added measures until SY 2010-2011. As a result, value added results from SY 2009-2010 are limited to classroom and school-wide measures, while teacher-level value added data are available only for SY 2010-2011.

During the years addressed by this study, the district was still weighing how best to estimate and make use of their new value added measures. As a result, none of the teacher value added results estimated for either school year were shared with school administrators prior to the end of SY 2010-2011. Thus, none of the personnel decisions or evaluative ratings reported in this study were directly informed or influenced by them.

The external vendor provided extensive technical details regarding their estimation procedure, which I summarize here. In brief, the district's value added measures are derived from the following formulation:

Dosage. Individual teachers (or classrooms, in SY 2009-2010), are matched to their students in each tested subject area. However, in cases of co-teaching or overlapping responsibility for a shared group of students, co-teaching teams are identified and student results are shared, with each student's dosage apportioned proportionally to different teachers. If two teachers teach the same subject to all of the same students, they each receive an identical value added measure.

Value added estimation. Individual teacher estimates are calculated based on the following formula:

$$(1)\ y_{ist} = \beta_1 y_{is(t-1)} + X_{it}\delta + \sum_{j-1}^{J} D_{ist}^{(j)} \tau_{js} + \varepsilon_{ist}$$

Where $y_{is(t-1)}$ is the student's pretest score expressed as a student z-score. X_{it} is a vector of other covariates – in this case student gender, poverty, race, and special education status. $D_{ist}^{(j)}$ is the student's dosage from teacher *j* in subject *s* in year *t*. τ_{js} is the true value added of teacher *j* in subject *s* averaged over the time period used in estimation (in this case, each of the individual school years was modeled separately). $y_{is(t-1)}$ and X_{it} are mean-centered around their respective sample means within individual grade and subject reference groups, in order to ensure that standard errors are relative to the average teacher.

The model utilizes a two-stage least squares estimation to correct for test measurement error, using each student's alternate subject pre-test score, *s'*, as an instrument.[4] The first stage equation is as follows:

$$(2)\ y_{is(t-1)} = \alpha_1 y_{is'(t-1)} + X_{it}\gamma + \sum_{j-1}^{J} D_{ist}^{(j)} \mu_{js} + \omega_{ist}$$

And obtains a predicted value for $\hat{y}_{is(t-1)}$, for each student's pretest score in subject *s*. The second stage of the procedure estimates a modified version of Equation 1 in which the predicted pre-test

score, $\hat{y}_{is(t-1)}$, is substituted for the actual pre-test score, $y_{is(t-1)}$. Standard errors in the second stage of estimation account for the clustering of outcomes by student. From the second stage of estimation, the estimated coefficient, $\hat{\tau}_{js}$, on the dosage variable for teacher *j* is the initial value added estimate for this teacher in subject *s*.

Precision adjustment. The initial value added estimates are then adjusted for precision by applying an Empirical Bayes estimation procedure that weights the initial estimate more heavily in accordance with its precision, and weights the group-wide average more heavily for less precise teacher estimates. The district first obtains an estimate of true variance in teacher value added estimates in subject *s*, $\hat{\sigma}_s^2$, via the iterative procedure outlined in Morris (1983). Then, the intermediate Empirical Bayes value added estimate for teacher *j* in subject *s*, denoted by $\tilde{\tau}_{js}$, is obtained as:

$$(3)\ \tilde{\tau}_{js} = \left(1 - \frac{d \times \hat{v}_{js}}{\hat{\sigma}_s^2}\right) \hat{\tau}_{js} + \left(\frac{d \times \hat{v}_{js}}{\hat{\sigma}_s^2 + \hat{v}_{js}}\right) A$$

Where *d* is a slight degrees-of-freedom adjustment specified in Morris (1983), and A is a weighted average of all initial value added estimates, with weights equal to $(\hat{\sigma}_s^2 + \hat{v}_{js})^{-1}$.

Conversion of metrics. The final value added estimate for each teacher is generated by subtracting the mean of all teacher value added estimates in the teacher's reference group from his or her own value added estimate, and dividing that value by the standard deviation of value added estimates in the reference group. For ease of interpretation, I convert this standardized performance metric to a percentile corresponding to a normal curve equivalent (NCE) distribution.

IV. Empirical Strategy

Overall Evaluation Scores

I address my first question of interest by generating an aggregate score to reflect a teacher's overall evaluation rating in each year. To do this, I standardize each of my individual indicators, and use a polychoric correlation matrix and principal component factor analysis across all of the

individual indicators to identify a single heavily-loaded factor. The resulting single factor explains 36% of the total variance in individual indicator ratings, and reflects a very high correlation (0.99) with a simple un-weighted mean score of the indicator ratings. I then standardize this factor.

I test this single factor measure as a predictor of the likelihood of each of my four personnel decisions of interest: dismissals, resignations, promotion to Assistant Principal of School Culture roles and promotion to Academic Assistant Principal and Principal roles in the same school year. For each of these outcomes, I run two separate logistic regression models, one with both the single evaluation factor and a vector of teacher demographic characteristics, and one with just a vector of teacher demographic characteristics. My demographic variables of interest include an indicator if a teacher's gender is female, a teacher's age in years in that school year, an indicator if a teacher is black,[5] and separate indicators for whether this is their first or second year of teaching in the district. I also include indicators for the individual school years 2008-2009 and 2009-2010.

$$(4)\ \ln\frac{\pi(p)}{1-\pi(p)} = \beta_0 + \beta_1 X_{it} + \beta_2 \delta_{it} + \gamma_t + \varepsilon_{it}$$

Here, the log likelihood of personnel decision p is a function of a vector X of teacher i's characteristics in year t, that teacher's rating δ in year t, and fixed effects for individual years, γ_t. I report model results as odds ratios corresponding to my dependent variables of interest.

Exploratory Factor Analysis

In order to examine whether a multiple-factor interpretation of the evaluation indicators better explains the data, I conduct an exploratory maximum likelihood factor analysis on teachers' indicator scores to identify patterns in the polychoric correlation matrix of the individual standardized ratings. Using the standard approach of consulting a scree plot and retaining factors with eigenvalues greater than 1.0, four underlying constructs emerge from the data that explain 81% of the cumulative variance.[6] To aid in the identification of patterns of loadings across factors, I use a varimax rotation.

One consequence of this rotation is that the rotated factors are uncorrelated with one another by construction, which affects how I interpret the results.

Across the 47 indicators, I identify and label four distinct dimensions, based on the pattern of high factor loadings detailed in Table 2. Each of these dimensions reflect a coherent interpretation, and is fairly consistent with the district's intended conceptual grouping of indicators in their evaluation rubric. Not all of the district-defined evaluation dimensions were identified as separate factors in the teacher ratings. However, indicators within each dimension were for the most part highly loaded onto a single factor. I standardize each of these four factors across my sample.

Multiple-Factor Prediction of Personnel Decisions

I test the predictive power of all four evaluation factors across several model variations. As in the single factor analysis, I predict the likelihood of each type of personnel decision in a separate model. I also, separately, predict the contemporaneous likelihood of each of the three anticipated personnel decisions – Expert Teacher, school leader within 1-2 years, and school leader within 3-5 years – using a sub-sample of teacher evaluation ratings from just SY 2009-2010. In each case, the model includes all four of the orthogonally rotated factors simultaneously as independent variables. Half of my model runs include demographic controls, and the other half do not. Models for personnel decisions include indicator variables for each of the individual school years in the sample, as in Equation 4 above. The only modification to Equation 4 is that here a teacher's rating δ in year t now refers to a vector of rating factors rather than a single rating score. Also, when predicting anticipated personnel decisions in SY 2009-2010 no year fixed effect is included.

Centered ratings. It is plausible that school administrators primarily operate within their local frame of reference when determining whether a teacher merits dismissal or promotion. If this is the case, it may be that an individual teacher's rating relative to other teachers at the same school and in the same school year is more relevant to administrator personnel decisions than an absolute rating. In

preliminary analyses, I tested each of the multi-factor logistic models using school-and-year-centered teacher ratings. However, since results using this specification were very similar to those from un-centered analyses, I do not include them in this version of the paper.

Predicting Individual Teacher Value Added Effectiveness

In order to test the relationship between teacher formative evaluations and student achievement gains in the same year, I use a linear regression model to predict each teacher's SY 2010-2011 value added NCE percentile as a function of first the single-factor and then the multiple factor ratings they received in the middle of that school year. I run separate models for the two subject areas of math and language arts. As before, I run each regression with and without demographic controls.

$$(5)\ \mu_{is}=\beta_0 + \beta_1 X_i + \beta_2 \delta_i + \omega_i + \varepsilon_i$$

Here, teacher *i*'s SY 2010-2011 value added percentile μ in subject *s* is a function of a vector of SY 2010-2011 evaluation rating factors δ and of teacher characteristics X. Because teachers were compared to external teachers from different external reference districts depending on the geographic location of their school, I also include a fixed effect ω corresponding to each teacher's region. Because very few teachers with value added data in SY 2010-2011 were promoted or dismissed, I do not investigate associations between value added and personnel decisions.

Predicting School-wide Value Added Effectiveness

Finally, I use school-wide evaluations of teachers to predict aggregate school-wide value added scores for the 10 district schools that participated in the formative evaluation program and had students in value added grades 4-8 in SY 2009-2010 or SY 2010-2011. In order to assess school-wide teacher contributions, I generate mean scores for each of the multiple-factor formative teacher evaluation ratings across each school, in each year. School-wide evaluation ratings include all rated teachers, not just those teaching in grades 4-8 in tested subjects.

In order to generate a measure of overall school value added effectiveness in each year, I leverage the classroom-level value added estimates from SY 2009-2010 and the teacher-level value added estimates from SY 2010-2011. I average the value added estimates of individual teacher or classroom value added in each school and year. I compute this mean for all teachers or classrooms at each school that participated in the mid-year evaluation system in that year, separately by subject area. Because the value added estimates are already shrunken according to the Empirical Bayes procedure, I weight each estimate equally in this calculation. For consistency and ease of interpretation, I then convert the final school-wide average scores to NCE percentiles.

An alternative approach to estimating school-wide value added would have been to directly compare adjusted student achievement gains in each school with that of comparison schools from the external benchmark districts. However, since the district was unable to directly share the external district data, I employ this simpler approach instead.

Using these measures, I run a linear regression to predict school-wide value added for each subject area, first with the school-wide averages all four of the factor evaluation ratings included, and then with each factor rating average on its own, in order to preserve my limited degrees of freedom.

$$(6)\ \partial_{ast}=\beta_0 + \beta_1\delta_{at} + \omega_a + \gamma_{at} + \varepsilon_i$$

Here, school a's value added percentile ∂ in subject *s* and year *t* is a function of a vector δ of one or more school-wide evaluation rating factors in the same year. I also include a fixed effect ω corresponding to school a's region, and a fixed effect γ for the 2009-2010 school year. In order to account for the clustering of data points within-schools and across years, I use a bootstrap to obtain standard errors, clustered at the school level.[7]

Across all of my school-level models, I individually investigated – but do not include – aggregated school-level demographic control variables. These included the average age of teachers at the school, average rates of 1st and 2nd year teachers at the school, the average rate of female teachers

at the school, and the average proportion of each teacher racial category of teachers at the school, in each year. I also investigated associations when controlling for school type (elementary versus middle schools). None of these school characteristics was substantially or significantly associated with school-wide value added.

V. Results

Descriptive Results

As shown in Figure 1, individual teachers' evaluation ratings reflect a relatively even distribution of high and low scores. Individual within-school rating distributions have a similar distribution. In addition, school-wide average standardized teacher ratings reflect substantial variation, with a range from -0.75 and 0.64, and a standard deviation of 0.40. These results indicate that administrators did not hold back from offering critical feedback to teachers or from evaluating their school staff overall as either low or high performing. This contrasts with typical public school subjective evaluations in current practice (Weisburg et al., 2009). The specificity of the indicators employed and the use of the evaluations for low stakes, formative mentoring may have encouraged frank appraisals. In the rest of this section, I summarize results pertinent to each of my research questions.

Do Overall Teacher Evaluation Ratings Predict Subsequent Dismissals and Promotions?

I find that administrators' overall formative teacher evaluation ratings are significant and substantial predictors of future personnel decisions, as illustrated in Table 3. A one standard deviation increase in a teacher's mid-year rating – using an overall evaluation measure – is associated with a 62% reduction in the likelihood of being dismissed, and a 24% reduction in the likelihood of resigning voluntarily. The same standard deviation increase in overall ratings predicts a seven-fold increase in the likelihood of promotion to a school culture leadership role, and a more than three-fold increase in the likelihood of promotion to a school academic leadership role. Results for both

dismissals and promotions are highly significant, and explain a substantial portion of the log likelihood of personnel outcomes.

Some demographic variables are also associated with particular personnel decisions. Teachers are substantially less likely to voluntarily resign after their first year in the district, perhaps opting to give the job more time. 1st year teachers are somewhat more likely to be dismissed, but only in a specification that does not control for evaluation ratings. Older teachers are less likely to resign and are more likely to be dismissed. Finally, black teachers are much more likely to be promoted to school culture leadership roles. Racial homophily with the majority of students at the school may be perceived to be a positive trait for that role's student and parent-facing responsibilities, or race may simply be correlated with other unobserved teacher skills or affinities that administrators believe to be important to the role.

Are There Distinct Factors Within Evaluative Ratings that Reflect Different Aspects of Teacher Performance?

As described in Section III, I identify four coherent and distinct factors that offer additional differentiation within the evaluative ratings provided by administrators. Table 2 summarizes the factor loadings of individual indicators onto the four factors. Based on the patterns of loadings, I label and briefly interpret each factor here, in sequence.

"Student Engagement and Behavior" appears to primarily reflect a teacher's observed classroom interactions with students. Demonstrated classroom management and motivational skills and a positive response by students are central. This factor also reflects how well the teacher reinforces school-wide cultural and behavioral norms.

"Instructional Specifics" reflects a teacher's skill at instructional execution as observed in classroom practice. It emphasizes specific district priorities related to instruction, including a high quantity of scaffolded practice by students. It also reflects a teacher's knowledge of curriculum, as reflected in both classroom practice and long term instructional planning.

"Personal Organization and Planning" appears to capture a teacher's basic organizational and planning skills. Some of the traits it includes may be observed in classroom practice, but most would be observed in other interactions with the teacher. The ability to create daily and long term instructional plans, as well as systematic tracking and use of student achievement data are emphasized. Interpersonal interactions with other staff and teachers' attention to continually improving their practice are also relevant.

"Parent and Student Relationships" corresponds to a teacher's management of relationships with student's parents, including their engagement on the topics of students' goals and progress. This factor also reflects a teacher's development of relationships with students outside of the classroom. Indicators in this factor would for the most part be observed outside of teachers' classroom practice.

Are Different Types of Personnel Decisions Predicted by Different Factors from the Evaluative Ratings?

Collectively, the four distinct factor ratings of a teacher explain substantially more variation in subsequent personnel decisions about that teacher than a single principal component factor. Moreover, the factors that administrators weigh in their decisions vary in accordance with the particular personnel decision in question. Table 4 provides a summary of model runs using all four orthogonally-rotated evaluation factors, in lieu of a single principal factor.

A teacher's rating in the category of Student Engagement and Behavior appears to be a major consideration in dismissal decisions, as well as in either type of school leadership promotion. A teacher who scores one standard deviation higher in this factor is less than half as likely to be dismissed, and two and half times as likely to be promoted to academic school leadership roles. In addition, the teacher is more than eight times as likely to be promoted to a school culture leadership role. Administrators' judgments about a teacher's ability to positively engage with students in their classroom and to reinforce the school-wide culture appear to be important across a variety of personnel decisions.

A high rating in Personal Organization and Planning is another major factor associated with a reduced likelihood of teacher dismissal. A one standard deviation increase in Personal Organization and Planning corresponds to a 46% reduction in the likelihood of dismissal. In addition, this factor predicts a smaller, but significant reduced likelihood of a teacher's voluntary resignation. Personal Organization and Planning also predicts decisions to promote teachers to academic leadership roles, but to a lesser degree than the factor for Student Engagement and Behavior. Administrators appear to value planning and organizational skills highly in their determination of whether to dismiss a teacher. They may also consider those skills relevant to some school leadership roles.

The factor reflecting Parent and Student Relationships strongly predicts promotion to school culture leadership roles, but is not a significant consideration in other personnel decisions. A one standard deviation increase in this factor rating corresponds to a nearly six-fold increase in the likelihood of promotion to school culture leadership. It appears that administrators consider this factor important for some teachers in specific school leadership roles, but value it less when considering teacher dismissals or promotions to academic leadership.

Instructional Specifics is not as predictive of personnel decisions as the other three factors, although it is a marginally significant predictor of the likelihood of promotion to either school culture or academic leadership roles, with odds ratios of 1.939 and 1.553, respectively.

Are Different Types of Anticipated Personnel Decisions Associated with Different Evaluation Factors?

Administrators' internal assessment of teachers' potential for future promotion are summarized in Table 5. In general, the results for anticipated teacher promotions to school leadership roles reflect the same priorities as in administrators' enacted promotion decisions, but with reduced effects sizes in terms of ratings on Student Engagement and Behavior. Interestingly, effect sizes associated with Personal Organization and Planning are about as large for teachers identified as

potential school leaders in either 1-2 or 3-5 years as they are for those who are actually promoted. This foundational skill set may be considered an important indicator of future potential.

Administrator's identification of potential "Expert Teachers" corresponds to high teacher ratings in Student Engagement and Behavior and in Instructional Specifics. They were twice as likely to identify a teacher as a potential Expert Teacher if that teacher scored one standard deviation higher in Instructional Specifics. A one standard deviation increase in Student Achievement and Behavior predicted nearly triple (2.815) the likelihood of Expert Teacher identification. Both of these factors appear relevant to administrators' evaluation of classroom teaching skills. It is notable that Instructional Specifics was a predictor of Expert Teacher identification, but not of reduced likelihood of dismissal, which suggests that this factor may reflect advanced skills that are not a central focus of formative feedback provided to struggling teachers.

Do Either Overall Ratings or Specific Factors Predict Teacher Value Added Performance in the Same Year?

Administrators' overall mid-year evaluation ratings of teachers do not predict student achievement gains in the same school year, but some individual factors do. Table 6 details associations between teachers' characteristics, their evaluation ratings,[8] and their value added NCE percentiles in the same school year.

When evaluative ratings are modeled as four separate factors, I find that Instructional Specifics is a significant positive predictor of teacher value added in language arts. A one standard deviation increase in teacher ratings on this factor predicts a 12% increase in a teacher's value added NCE percentile. The instructional techniques emphasized in teacher evaluations may be relevant to improving student achievement in this area. Alternatively, implementation of those instructional techniques may simply be correlated with a successful learning environment for students.

Higher teacher ratings in Parent and Student Relationships predict reduced student achievement gains in math and in language arts, in most model specifications. It may be that attending too much

to Parent and Student Relationships detracts from teachers' effectiveness in the classroom. Alternately, greater observed engagement with students' parents may reflect a teacher's response to already struggling students.

Available teacher demographics show limited ability to predict value added. Students of older teachers in this district have reduced learning gains in math, which may partially explain why older teachers are more likely to be dismissed. Teachers who are newer to the district appear less effective in math, but only when controlling for formative evaluation ratings. The inconsistent association between experience in the district and teacher value added is not surprising, since the district primarily hires teachers with more than 2 years of prior teaching experience.

Overall, the modest associations between the evaluations and same-year value added do not necessarily suggest that administrators ignore or fail to identify teachers' value added performance when determining promotions and dismissals. The low-stakes formative feedback they provide teachers at mid-year is unlikely to fully capture every dimension of teaching that administrators may consider in their ultimate personnel management decisions.

Do School-wide Teacher Evaluation Ratings Predict School-wide Value Added in the Same Year?

Most school-wide factor ratings do not predict student achievement gains. However, I find that the school-wide average teacher rating for Student Engagement and Behavior predicts significantly higher school-wide average student achievement gains in math. Table 7 details how the average evaluative rating of all the teachers in a school predict that school's average value added performance in the same year in the tested subjects of math and language arts.

It is important to note that my sample size is small in these analyses, with just 10 participating schools with value added results in SY 2009-2010 and 9 schools in SY 2010-2011. As such, most estimates in these tables are insignificant and the standard errors reflect a high degree of noise. However, the point estimate on the association between school-wide Student Engagement and

Behavior ratings and school-wide math value added is very large (35.475 NCE percentage points), and is significant with a t-table derived p-value of 0.020. Separate investigations of this effect in each school year (not shown) indicate comparable effect sizes across years, and statistically significant results in SY 2010-2011 alone.

For this district, the cumulative contributions of the entire teacher team towards engaging students and managing student behavior predict school-wide student achievement gains in math. This association may help to explain why so many administrator personnel decisions are predicted by the Student Engagement and Behavior factor, more-so than teachers' ratings in other factors. It is possible that administrators value teacher performance in this area because they believe it contributes to positive school-wide results. It may also be that administrators tend to rate their teachers higher in this area in schools that have previously established cultures of positive student engagement and behavior. That said, even within schools-and-years characterized by a high average score in this factor, individual teacher ratings vary substantially and are not uniformly high.

It is unclear the extent to which this school-wide association may reflect peer teacher effects, versus a combination of peer and individual effects.[9] However, it is notable that, as shown in Table 6, an individual teacher's rating in Student Engagement and Behavior does not by itself predict significant differential student math achievement gains in his or her own classroom.

VI. Conclusions and Discussion

This study offers new insights into the diversity of teacher contributions that local school administrators value in their management of school staff. I find that administrators consider a range of teacher practices when assessing teacher quality, and that their formative evaluations predict their future personnel decisions. Moreover, administrators weigh different, distinct evaluative criteria when staffing different teacher roles at their schools. The results suggest the importance of accounting for multiple aspects of teacher performance in evaluation systems that are meant to

inform multiple types of personnel decisions. Attention to multiple measures of performance may be particularly relevant in schools where teacher roles are more differentiated and reflect a mix of leadership, mentoring, family relations, and instructional responsibilities.

Some of the criteria valued by administrators in this district could be observed in teachers' daily classroom practice, while others reflect teachers' planning, their interactions with peers, or their interactions with students and families that are more likely to be observed outside of the classroom. This suggests that evaluation systems may be more accurate if at least some of the multiple measures used to evaluate teachers address their contributions and competencies outside of instructional execution. In particular, teacher dismissals appear to be informed not only by instructional execution, but by observations of teachers' pre-planning and organizational skills.

Associations between observable teacher demographic characteristics, such as race or age, and personnel decisions should be interpreted with caution. In contrast to the mid-year evaluation ratings, there is no direct evidence demonstrating that administrators actively considered demographic traits in their judgments about teachers' performance or contributions. Instead, the demographic trends that I observe suggest potential areas of future inquiry for research examining administrators' personnel management priorities.

I also identify an association between a teacher evaluation factor that administrators valued highly across all of their personnel decisions – Student Engagement and Behavior – and school-wide average student achievement gains in math. Schools characterized by strong teacher ratings in this area were more effective at raising student achievement than those that were not. The limited sample size renders this finding somewhat tenuous. Nevertheless, the results offer suggestive evidence that evaluation systems may be more effective if they weigh teachers' coordinated efforts in addition to their individual expertise.

The evaluation criteria used in the formative teacher evaluations were of only modest value in predicting teachers' individual value added performance in the same school year. In part, this may reflect the limitations of relying upon just a single year of linked teacher-and-student data to assess individual teachers' performance. In addition, it is unclear the extent to which school administrators in this district actively sought to provide teachers with formative feedback about their individual competency in raising student achievement. Nevertheless, the results suggest that if they intend to do so, they might improve their accuracy by referencing objective teacher value added measures. The fact that the district evaluation rubric instead considered only absolute student performance relative to fixed achievement goals likely reflected a motivational aspect of the formative evaluations. However, this focus on absolute levels of student achievement may have biased the teacher evaluations unhelpfully (Jacob and Lefgren, 2008).

Finally, it is important to note that the district in this study represents something of an outlier from other public schools. This is demonstrably the case in terms of enacted school personnel management policies and evaluation practices. Thus, the specific administrator priorities that I identify may not be replicated in other districts. Different schools will vary in the teacher roles and promotions that they designate. They may also vary in terms of the shared professional norms and practices that are considered essential to teaching. For example, not all schools will separate assistant principal roles into "academic" and "culture" categories. Similarly, conceptions of what constitutes expert instruction will vary across school contexts, subject areas, and grade levels.

Moreover, it is also plausible that administrators in this district possess unique management skills. At a minimum, they have experience managing a higher volume of teacher dismissals than is typical in most public school districts. In addition, as leaders within a demonstrably high-performing school district, they may be unusually skilled at identifying important teacher contributions or

providing formative feedback. Districts who attempt to implement similar evaluation or personnel management practices may not experience the same results, at least not initially.

Nevertheless, this study provides insights about teacher evaluation that are relevant to a broad range of contexts. The complexity involved in assessing teacher contributions in this district is likely to be common across a variety of school settings. For example, even schools with little formal differentiation in teacher roles may distribute important responsibilities across informal teacher roles, such as subject and grade level chairs or co-teacher mentors. Similarly, diverse teacher contributions, such as supporting school-wide effectiveness or peer learning, are likely to be important across a variety of school contexts (Jackson et. al, 2009; Ronfeldt et. al, 2011). Evaluation systems that account for the potential complexity and diversity of teachers' contributions by including more holistic and flexible measures of their performance may be more responsive and ultimately more effective at improving student outcomes. Local administrators are likely to play key roles in this regard, and their perspectives may be critical to the evolution and improvement of emerging teacher evaluation systems.

Although it is unique, this district usefully demonstrates the practices of a group of highly autonomous local administrators in high performing schools. They were afforded a substantial degree of discretion in their personnel management decisions. To support their work, they leveraged a standards-based evaluation instrument that is similar to other emerging teacher evaluation measures. In these respects, their experience provides an important test case for other public school districts as they embark on evaluation system reforms. Such reforms will likely result in greater discretion and responsibility for the local school administrators who manage their implementation. This district's example suggests some of the tradeoffs that those administrators may face, and how they may respond to them in practice.

Notes

[1]District schools have been shown to be demonstrably highly effective at raising student achievement, in comparison to lottery-randomized comparison students attending nearby schools.

[2]In order to preserve district anonymity, I include only essential organizational details. Interested parties may contact me with clarifying questions regarding additional context pertinent to the study.

[3]To help preserve district anonymity, I do not include a copy of the district's evaluation form.

[4]Use of the alternate subject pre-test score as an instrument to eliminate measurement error could also introduce bias depending on the relationship between students' scores in either subject. The district examined value added measures that instead control for both pre-tests but do not instrument to eliminate measurement error. This alternate specification yielded teacher value added rankings that are highly correlated (0.86) with their chosen method.

[5]I also explored the use of teacher controls for Asian, Hispanic, and "other" races, but these were neither directionally nor significantly associated with any outcomes of interest.

[6]Eigenvalues of these four factors were 17.668, 2.218, 1.511, and 1.036. The next highest was 0.938 and an investigation of the scree plot and its factor loadings did not support its inclusion as a coherent, distinct factor.

[7]For all model runs, I perform 1000 iterations. I use a seed # of 1000 to enable replication of results.

[8]I also separately investigated whether the particular formative evaluation indicator that assesses teachers' mid-year student achievement levels relative to fixed achievement goals predicts teacher value added in the same year. There was no significant association between the two.

[9]I investigated alternative model specifications to distinguish between the effects of same-subject and other-subject peer evaluative ratings on school-wide value added, with inconclusive results.

Tables and Figures

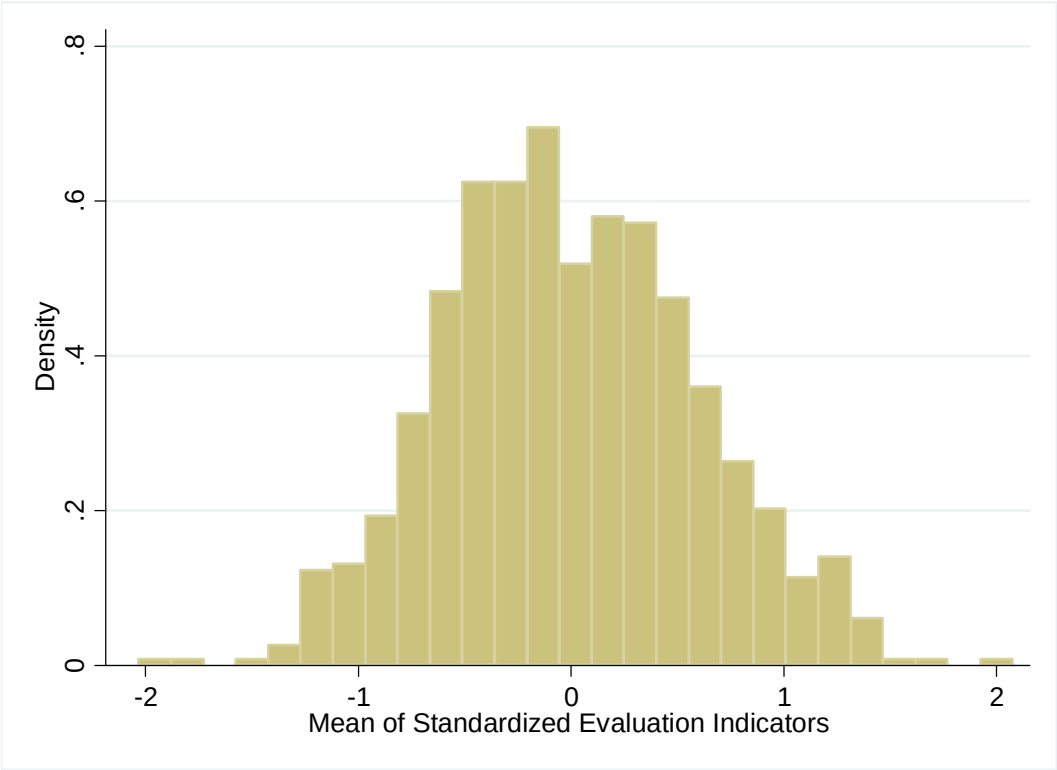

FIGURE 1. *Distribution of the Average of Individual Teachers' Standardized Indicator Ratings*

TABLE 1
Descriptive Statistics for District Teachers and Personnel Decisions, by School Year

	SY 2008-2009	SY 2009-2010	SY 2010-2011	All Years
# of Teachers Evaluations	178	263	306	747
Average Within-School Evaluation Rate	64%	78%	83%	77%
% Promoted to Academic School Leadership	3.9%	2.7%	0.7%	2.1%
% Promoted to School Culture Leadership	2.8%	1.5%	0.7%	1.5%
% Resigned	11.2%	9.1%	12.7%	11.1%
% Dismissed	7.9%	8.7%	5.6%	7.2%
% with Expert Teacher potential		14.4%		
% with Leadership potential, next 1-2 years		4.6%		
% with Leadership potential, next 3-5 years		8.4%		
% Female	77.5%	74.5%	71.2%	73.9%
% White	65.2%	64.3%	66.3%	65.3%
% Black	18.5%	16.3%	13.1%	15.5%
% Hispanic	8.4%	8.4%	10.1%	9.1%
% Asian	3.4%	4.9%	4.9%	4.6%
% Other/ Unknown	4.5%	6.1%	5.6%	5.5%

TABLE 2

Conceptual Structure and Factor Loadings of Mid-Year Formative Teacher Evaluation Indicators

Evaluation Dimension	Evaluation Indicator	Factors			
		Student Engagement and Behavior	Instructional Specifics	Personal Organization and Planning	Parent & Student Relationships
Achievement	Achievement relative to goals				
Character	Students respectful	0.77			
	Students enthusiastic	0.77			
	Students do their best	0.52			
	Students' citizenship	0.62			
	Students present/prepared	0.54			
Instruction	Clear goals for each lesson			0.53	
	Daily assessment			0.45	
	Accurate content		0.56		
	Well-planned lesson		0.64		
	Clear lesson sequence				
	Guided practice		0.61		
	Checks for understanding		0.57		
	Independent practice		0.49		
	Support during ind. practice				
	Student work time		0.57		
	Quality responses		0.41		
	Quality questions		0.43		
	Differentiation		0.41		
Classroom Culture	All Students on-task	0.67			
	Engagement strategies	0.62			
	Classroom routines	0.75			
	High behavioral standards	0.72			
	Positive classroom environment	0.62			
	Positive student interactions	0.54			
	Character building	0.53			
	Tie character to lessons	0.44			
	Neat / orderly classroom				
	Support school culture system	0.44			
	Proper use of incentives	0.49			
Systems and Planning	Goal-setting			0.48	
	Investing students in goals				
	Knowledge of curriculum		0.50		
	Year-long instructional plan		0.44	0.49	
	Unit plans		0.44	0.48	
	Lesson plans			0.53	
	Weekly/informal data use			0.48	
	Organized data tracking			0.44	
	Periodic/formal data use			0.52	
Student and Family Relationships	Cares about students	0.41			0.40
	Relationships outside of class				0.50
	Relationships with families				0.73
	Sharing goals with parents				0.53
	Communication with parents				0.66
Personal Effectiveness	Constantly learning			0.41	
	Organized			0.52	
	Attendance				
	Communication with peers			0.41	

Note: Indicators that are not highly loaded (>0.40) on any single factor are left blank

TABLE 3

Predicting the Likelihood of Teacher Dismissals, Resignations, and Promotions with Teacher Characteristics and Overall Evaluation Ratings (Odds Ratios)

	Teacher Characteristics				Teacher Characteristics and Evaluation Ratings			
	Dismissed	Resigned	Promoted to AP of Culture	Promoted to Academic AP or Principal	Dismissed	Resigned	Promoted To AP of Culture	Promoted to Academic AP or Principal
1st Year in the District	2.174~	0.258***	0.413		0.874	0.201***	2.558	
	(0.863)	(0.087)	(0.396)		(0.373)	(0.072)	(2.849)	
2nd Year in the District	1.782	0.848	1.630		1.358	0.790	3.348	
	(0.739)	(0.241)	(1.200)		(0.570)	(0.226)	(2.739)	
Female	0.612	1.175	2.797	0.745	0.621	1.180	2.762	0.744
	(0.187)	(0.338)	(2.977)	(0.411)	(0.198)	(0.341)	(3.028)	(0.428)
Black	1.675	1.200	3.532*	0.325	1.314	1.113	8.336**	0.388
	(0.600)	(0.373)	(2.240)	(0.339)	(0.494)	(0.349)	(6.345)	(0.411)
Age	1.094***	0.925*	0.966	1.012	1.122***	0.934*	0.925	0.949
	(0.023)	(0.030)	(0.073)	(0.049)	(0.026)	(0.031)	(0.084)	(0.058)
Indicator for SY 2008-2009	1.377	0.946	3.889	6.703*	1.633	0.989	3.541	8.538*
	(0.535)	(0.287)	(3.344)	(5.433)	(0.657)	(0.301)	(3.260)	(7.178)
Indicator for SY 2009-2010	1.710	0.617~	2.033	4.328	1.842~	0.617~	2.813	5.619*
	(0.582)	(0.172)	(1.780)	(3.494)	(0.644)	(0.173)	(2.615)	(4.685)
Evaluation Rating, Single Factor					0.377***	0.756*	7.078***	3.395***
					(0.071)	(0.103)	(3.472)	(1.033)
Number of Teacher-Ratings	747	747	747	747	747	747	747	747
Psuedo R-Squared	5.907%	5.511%	11.471%	5.784%	13.721%	6.337%	30.808%	17.739%

Note: AP = Assistant Principal. As no 1st-year teachers were promoted to Academic AP or Principal roles, experience controls are omitted in those models.
~p<.1 *p < .05, **p < .01, ***p < .001.

TABLE 4

Predicting the Likelihood of Teacher Dismissals, Resignations, and Promotions with Multiple Evaluation Factors (Odds Ratios)

	Dismissed		Resigned		Promoted To AP of Culture		Promoted to Academic AP or Principal	
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Demographic Controls		Y		Y		Y		Y
Student Engagement and Behavior	0.443***	0.373***	0.931	0.788~	3.518***	8.386***	2.628***	2.664***
	(0.076)	(0.071)	(0.110)	(0.103)	(1.233)	(4.793)	(0.753)	(0.779)
Instructional Specifics	1.171	0.935	1.019	0.992	1.354	1.939~	1.482	1.553~
	(0.185)	(0.157)	(0.122)	(0.130)	(0.410)	(0.755)	(0.359)	(0.405)
Personal Organization and Planning	0.533***	0.541***	0.809~	0.764*	0.943	1.204	1.916*	1.834*
	(0.081)	(0.088)	(0.096)	(0.096)	(0.314)	(0.503)	(0.510)	(0.501)
Parent and Student Relationships	1.013	0.954	1.151	1.044	3.857***	5.852***	0.970	0.994
	(0.147)	(0.144)	(0.133)	(0.129)	(1.468)	(2.838)	(0.272)	(0.286)
Number of Teacher-Ratings	747	747	747	747	747	747	747	747
Psuedo R-Squared	11.732%	17.558%	1.275%	7.100%	32.291%	45.684%	19.072%	20.022%

Note: AP = Assistant Principal. Demographic controls include years of experience in the district, age, race, and gender. ~p<.1 *p < .05, **p < .01, ***p < .001.

TABLE 5

Predicting the Likelihood of Teacher Identification for Future Promotions with Multiple Evaluation Factors (Odds Ratios)

	"Master Teacher" Potential		School Leadership Potential within 1-2 Years		School Leadership Potential within 3-5 years	
Demographic Controls		Y		Y		Y
Student Engagement and Behavior	2.641***	2.492***	2.358*	2.073*	1.429	1.721*
	(0.574)	(0.589)	(0.787)	(0.742)	(0.334)	(0.454)
Instructional Specifics	2.050***	1.943**	1.360	1.198	1.332	1.453
	(0.409)	(0.406)	(0.396)	(0.363)	(0.308)	(0.363)
Personal Organization and Planning	1.393~	1.352	1.998*	1.978*	1.780*	1.830*
	(0.264)	(0.271)	(0.599)	(0.626)	(0.425)	(0.476)
Parent and Student Relationships	1.394	1.329	1.017	0.893	1.516~	1.789
	(0.298)	(0.298)	(0.354)	(0.353)	(0.378)	(0.483)
Number of Teacher-Ratings	253	253	253	253	253	253
Psuedo R-Squared	20.034%	20.793%	14.149%	17.379%	8.836%	12.813%

Note: Demographic controls include years of experience in the district, age, race, and gender. ~p<.1 *p < .05, **p < .01, ***p < .001.

TABLE 6

Predicting SY 2010-2011 Teacher Value Added NCE Percentiles with Teacher Characteristics and Evaluation Factors

	Math Value Added				Language Arts Value Added			
	Single Factor		All Factors		Single Factor		All Factors	
Region Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y
1st Year in the District		-16.910		-21.818*		4.970		7.479
		(10.347)		(10.219)		(12.422)		(11.577)
2nd Year in the District		-14.147		-24.799*		-5.845		-0.628
		(11.452)		(12.014)		(11.560)		(10.749)
Female		5.233		9.543		11.242		11.019
		(7.017)		(7.021)		(8.917)		(8.081)
Black		0.872		6.910		10.413		11.594
		(10.948)		(12.310)		(11.968)		(10.785)
Age		-1.206*		-1.062~		-0.753		-0.091
		(0.535)		(0.528)		(0.896)		(0.833)
Evaluation Rating, Single Factor	4.255	0.327			-1.745	0.647		
	(3.278)	(3.935)			(3.907)	(4.638)		
Student Engagement and Behavior			4.840	2.577			-3.017	-2.081
			(3.244)	(3.149)			(4.129)	(4.435)
Instructional Specifics			2.887	-0.471			11.948**	12.450**
			(3.388)	(3.853)			(4.001)	(4.232)
Personal Organization and Planning			5.124	5.463			1.376	0.795
			(3.452)	(4.178)			(3.733)	(3.821)
Parent and Student Relationships			-6.044	-10.528*			-10.485*	-9.596*
			(4.071)	(4.261)			(3.909)	(4.215)
Number of Teachers	38	38	38	38	53	53	53	53
Adjusted R-Squared	<0%	5.571%	1.883%	17.588%	<0%	<0%	17.776%	17.330%

Note: Teacher value added estimates are reported as normal curve equivalent percentiles, ranked relative to the mean and standard deviation of external-district teacher estimates in the same grade and subject. Sample includes 91 (out of a total of 120) district teachers with value added data for whom evaluation ratings were also available in SY 2010-2011. ~p<.1 *p < .05, **p < .01, ***p < .001.

TABLE 7

Predicting School-wide Value Added NCE Percentiles with the School-wide Average of Teachers' Mid-Year Evaluations

	Math					Language Arts				
	All Factors	Student Engagement and Behavior	Instr. Specifics	Org. and Planning	Parent and Student Relations	All Factors	Student Engagement and Behavior	Instr. Specifics	Org. and Planning	Parent and Student Relations
Region Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Year Fixed Effect	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Student Engagement and Behavior	39.23~ (23.582)	35.475* (15.289)				3.238 (17.744)	-1.653 (13.103)			
Instructional Specifics	-8.900 (23.963)		-0.305 (15.62)			-9.812 (11.467)		-9.064 (8.031)		
Personal Org. and Planning	-2.880 (20.547)			11.263 (12.92)		-1.959 (15.371)			-0.543 (12.661)	
Parent and Student Relationships	12.615 (31.764)				13.467 (20.033)	0.243 (17.927)				-1.23 (13.108)
Number of Schools	19	19	19	19	19	19	19	19	19	19
Adjusted R-Squared	29.794%	37.350%	11.279%	16.734%	16.622%	13.428%	23.635%	30.326%	23.548%	23.611%

Note: Sample of schools with students in value added grades (4-8). School-wide averages of teacher evaluation ratings calculated as the average of all individual teacher ratings, including teachers without value added scores. School-wide average value added calculated as the average of teachers' (SY 2010-2011) or classrooms' (SY 2009-2010) value added percentiles, including teachers who were not evaluated. Bootstrap standard errors clustered at the school level. ~p<.1 *p < .05, **p < .01, ***p < .001.

References

Ballou, D. (2000). *Teacher Contracts in Massachusetts*. Boston, MA: Pioneer Institute for Public Policy.

Balu, R., Beteille, T., and Loeb, S. (2010). "Examining teacher turnover: The role of school leadership." *Politique Americaine* 15:55-79.

Bill & Melinda Gates Foundation. 2012. "Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains." MET Project Research Paper.

Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J (2010, July 20). "Teacher layoffs: An empirical illustration of seniority v. measures of effectiveness." CALDER working paper.

Chingos, M., & West, M. R. (2011). Promotion and reassignment in public school districts: How do schools respond to differences in teacher effectiveness? *Economics of Education Review*, *30*(3), 419–433.

Gallagher, H. A. (2004) Vaughn Elementary's innovative teacher evaluation system: are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education* 79 (4), 79–107.

Goldhaber, D., and Theobald, R.. (2011) "Managing the teacher workforce in austere times: The implications of teacher layoffs." Seattle, WA: University of Washington Center for Education Data and Research, available at http://www.cedr.us/papers/working/CEDR%20WP%202010-7%20Teacher%20Layoffs%203-15-2011.pdf

Grossman, P., & Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., Lankford, H. (2010). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added. Working Paper No. 16015. Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w16015

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first grade classroom make a difference for children at risk of school failure? *Child Development, 76,* 949–967.

Harris, D. N., & Sass, T. R. (2009). *What makes for a good teacher and who can tell?* (CALDER Working Paper 30). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. http://www.caldercenter.org/upload/CALDER-Working-Paper-30_FINAL.pdf

Heneman, H.G. III., Milanowski, A., Kimball, S. M., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay.* CPRE Policy Brief, RB-45. Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Hess, F. M., and Loup, C. (2008). *The leadership limbo: Teacher labor agreements in America's fifty largest school districts.* Washington, DC: Thomas B. Fordham Institute.

Holtzapple, E. (2003) "Criterion-related validity evidence for a standards-based teacher evaluation system." *Journal of Personnel Evaluation in Education*, 17(3): 207- 219.

Jackson, C. K, and Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics* 1:85-108.

Jacob, B. A. (2010). Do principals fire the worst teachers? NBER Working Paper #15715.

Jacob, B. and Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*. 26(1), 101-36.

Jacob, B. A. and Walsh, E. (2010) What's in a rating? *Economics of Education Review*, 30, 434-448.

Kane, T. J.,Taylor E. S., Tyler J. H., and Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data, *Journal of Human Resources*, 587-613.

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

Milanowski, A. T. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53.

Morris, C. (1983) "Parametric empirical bayes inference: Theory and applications." *Journal of the American Statistical Association*, vol. 78, no. 381, pp. 47-55.

Price, M. (2009). *Teacher union contracts and high school reform.* Seattle, WA. Center on Reinventing Public Education.

Rockoff, J. E. and Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: evidence from New York City. *Labour Economics* 18: 687.696

Rockoff, J. E., Staiger D. O., Kane, T. K., and Taylor, E. S. (2011). Information and employee evaluation: evidence from a randomized intervention in public schools. *American Economic Review.*

Ronfeldt, M., Lankford, H., Loeb, S., Wyckoff, J. (2011). How teacher turnover harms student achievement. NBER Working Paper No. 17176.

Taylor, E.S., & Tyler, J.H. (2011). "The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers." National Bureau of Economic Research working paper no. 16877.

Weisberg, D., Sexton, S., Mulhern, J., and Keeling, D. (2009). The widget effect. *Education Digest,* 75(2), 31–35.