# The Sensitivity of Teacher Performance Ratings to the Design of Teacher Evaluation Systems

Matthew P. Steinberg
*University of Pennsylvania*

Matthew A. Kraft
*Brown University*

April, 2016
Updated: June, 2017

## Abstract

In recent years, states and districts have responded to federal incentives by instituting major reforms to their teacher evaluation systems. The passage of the Every Student Succeeds Act in 2015 now provides policymakers with even greater autonomy to redesign existing evaluation systems. Yet, little evidence exists to inform decisions about two key system design features – teacher performance measure weights and performance ratings thresholds. Using data from the Measures of Effective Teaching study, we conduct simulation-based analyses that illustrate the critical role that performance measure weights and ratings thresholds play in determining teachers' summative evaluation ratings and the distribution of teacher proficiency rates. These findings offer insights to policymakers and administrators as they refine and possibly remake teacher evaluation systems.

Suggested Citation:

Steinberg, M.P., & Kraft, M.A. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher*. *Educational Researcher, 46*(7), 378–396.

Link to Publisher's Version:

http://journals.sagepub.com/doi/full/10.3102/0013189X17726752

The authors thank Allison Atteberry, Cory Koedel, and Eric Taylor for feedback on previous versions of this paper, and participants at the 2016 Association for Education Finance and Policy conference for valuable comments and discussions. The authors thank Filippo Bulgarelli and Mariela Mannion for excellent research assistance, and Jennifer Moore for editorial assistance. Steinberg (corresponding author) may be contacted at steima@upenn.edu; Kraft may be contacted at mkraft@brown.edu.

## Introduction

In recent years, policy reforms at the federal, state, and local levels have dramatically changed the ways that educators are evaluated (Donaldson & Papay, 2015). These reforms, along with growing public scrutiny, arose from widespread recognition that traditional teacher evaluation systems neither differentiated among low- and high-performing teachers (Donaldson, 2009; Toch & Rothman, 2008; Tucker, 1997; Weisberg et al., 2009) nor provided teachers with meaningful feedback about their practice (Almy, 2011; Sartain, Stoelinga, & Brown, 2011; Sinnema & Robinson, 2007; Stronge & Tucker, 2003). By the 2015–2016 school year, 88% of both states and the largest 25 districts and the District of Columbia had revised and implemented new teacher evaluation systems (Steinberg & Donaldson, 2016).

Traditional systems of teacher evaluation tended to be perfunctory exercises, relying on a single measure of teacher performance (typically a cursory observation of classroom practice), binary summative ratings (i.e., proficient or not), and few if any consequences tied to teachers' summative ratings (Weisberg et al., 2009). In contrast, teachers' evaluation ratings under newly implemented evaluation systems have become increasingly high-stakes for both individual teachers and for districts as a whole. Policymakers and the public are increasingly asking districts to reconcile teachers' evaluation ratings with the performance of their students. This is in light of evidence that, under both traditional evaluation systems and many newly implemented systems, nearly all teachers continue to be rated professionally proficient (Anderson, 2013; Steinberg & Sartain, 2015; Kraft & Gilmour, in press).[1]

Efforts to reform teacher evaluation systems have been focused on three primary system design features: the incorporation of multiple measures of teacher performance; the use of multiple performance ratings categories; and the creation of professional support and incentive structures tied to teachers' ratings. District policymakers have been deeply engaged in decisions about which performance metrics should be incorporated into their evaluation systems, including test-based performance measures such as value-added measures (VAMs) or student growth percentiles (SGPs), as well as rubric-based observation ratings of a teacher's instructional practice. Further, nearly all new systems have expanded the range of performance ratings to include at least four categories defining a teacher's summative performance. Teachers who receive low ratings – typically the bottom two ratings categories – are now overwhelmingly required to participate in additional targeted professional development and are increasingly at risk of being terminated or non-renewed during the tenure review process (Steinberg & Donaldson, 2016).[2] Teachers with exemplary ratings may be rewarded with merit pay or promoted to new positions on a career ladder (Donaldson & Papay, 2015; Steinberg & Donaldson, 2016).

Research on teacher evaluation reforms has mirrored these patterns. Most existing studies focus on the reliability and validity of performance measures—VAMs (e.g., Chetty, Friedman & Rockoff, 2014; Kane, McCaffrey, Miller, & Stiager, 2013), classroom observation rubrics (e.g., Garrett & Steinberg, 2015; Hill, Charalambous, & Kraft, 2012; Kane & Staiger, 2012) and student surveys (e.g., Kane & Cantrell, 2010; Wallace, Kelcey, & Ruzek, 2016). A related line of research evaluates how these new

high-stakes systems affect teacher performance, student achievement and teacher turnover (Cullen, Koedel & Parsons, 2016; Dee & Wyckoff, 2015; Steinberg & Sartain, 2015; Sartain & Steinberg, 2016). Even practitioner-facing guidebooks and edited volumes have primarily focused on how to design evaluation systems to more reliably evaluate teachers and/or leverage the evaluation process to promote teacher development (Darling-Hammond, 2013; Grissom & Youngs, 2015; Kane, Kerr, & Pianta, 2014; Marzano & Toth, 2013).

With this paper, we illustrate the central role that two equally important system design features play in shaping teachers' summative evaluation ratings, but which have received far less policy and research attention: performance measure weights and summative evaluation ratings thresholds. In comparison to decisions about which measures to choose and how to design consequential incentives, state and local policymakers have almost no empirically-based evidence to inform their decision process about how to combine scores across multiple performance measures and then how to map these summative evaluation scores onto performance ratings categories. Informal conversations with administrators and researchers involved in the design process suggest that decisions about weights and performance ratings thresholds are often made through a somewhat arbitrary and iterative process, one which is shaped by political considerations in place of empirical evidence. As we demonstrate in this paper, these decisions can have important consequences for both individual teachers' ratings and the share of teachers deemed to be professionally proficient.

The passage of the Every Student Succeeds Act (ESSA) in December 2015 makes research that can inform the evaluation system design process more important now than ever before. ESSA has ushered in a new phase in the teacher evaluation reform movement by granting states and districts considerable autonomy to redesign and implement teacher evaluations systems independent of federal influence. Research that informs the evaluation system design process is especially important given the existence of what Richard Elmore (2002) termed the "capacity gap" in state departments of education – the gap between what they are expected to do and what they are staffed to accomplish (Le Floch, Boyle, & Therriault, 2008). Several recent studies have found that limited technical expertise in state departments of education has constrained their ability to fully attend to all important design features of teacher evaluation systems (Herlihy et al., 2014; McGuinn, 2012).

We address this need by conducting simulation-based analyses to examine how teachers' summative evaluation ratings are affected by the decisions district administrators make about the weights they assign to multiple performance measures and the ratings thresholds that they choose. Specifically, we investigate how the proportion of teachers deemed professionally proficient changes under different weighting and ratings thresholds schemes. We examine how teacher proficiency rates change when we vary the performance weights (holding the ratings thresholds scheme fixed), when we vary the ratings thresholds scheme (holding the performance weights fixed), and how these design decisions interact with one another (i.e., when we jointly vary performance weights and ratings thresholds). Our analyses also allow us to provide additional empirical insight into

how the properties of teacher evaluation measures – specifically, the mean, variance and cross-measure correlations – shape the distribution of teacher proficiency ratings under these different system design parameters.

It is straightforward to infer that teacher proficiency rates will improve as, for example, greater weight is given to performance measures with higher average scores and/or the minimum threshold required to receive a Proficient rating is set lower. Ours is the first paper, to our knowledge, to more precisely illustrate the degree to which marginal changes in the weights assigned to performance measures and the placement of ratings thresholds can shift the distribution of teacher ratings and, ultimately, affect the proportion of teachers deemed professionally proficient. Though our findings are not intended to provide specific recommendations about *what* weights and ratings to select – such decisions are fundamentally subject to local district priorities and preferences – they do offer important insights about *how* these decisions will affect the distribution of teacher performance ratings as policymakers and administrators continue to refine and possibly remake teacher evaluation systems.

We accomplish this by drawing on rich data collected by the Measures of Effective Teaching (MET) Project. The MET Project affords a unique opportunity to examine the sensitivity of teacher performance ratings. In particular, the MET data contain a wide range of performance measures that are common to more than 1,000 teachers and which are currently being incorporated into new teacher evaluation systems. We use these data to illustrate how the summative performance ratings for the *same* set of MET teachers change as we impose different evaluation design parameters based on

existing evaluation systems across a range of large, urban school districts. To do this, we first construct teacher evaluation scores from combinations of three performance measures found in many new teacher evaluation systems: scores from classroom observation rubrics, estimates of teachers' contributions to student achievement, and student survey responses capturing their perceptions of teacher performance in the classroom. We then combine these data with publicly available information on the performance ratings thresholds currently used across eight large and geographically diverse urban school districts. Together, these data allow us to conduct a range of simulation analyses that illustrate the consequences of different weighting regimes and ratings thresholds. While our analyses focus on teachers in tested grades and subjects, we also discuss how our findings relate to the evaluation ratings received by the majority of teachers who teach in non-tested grades/subjects.

We first describe the considerable variation across districts in both the weights they assign to different performance measures and the percent of available evaluation points required to earn a given summative evaluation rating. We then show how teachers can receive substantially different summative ratings, with the same underlying scores on individual performance measures, depending on how weights are assigned to individual performance measures and how summative performance scores map on to summative rating categories. Our findings also reveal the important role that the properties of teacher performance measures play in determining teacher proficiency rates. First, if all performance measures are assigned equal weight, then the measure with the highest cross-measure correlation will contribute the most to a teacher's summative evaluation

score. Second, teacher performance measures that are weakly correlated with the other measures will contribute less to a teacher's summative score than would be expected based on the weight that the evaluation system assigns to it. And third, teacher proficiency rates depend not just on the properties of performance measures, but also on the location of the proficiency threshold relative to the actual distribution of teachers' summative evaluation scores. In evaluation systems where the proficiency threshold is located near the center of the distribution of teachers' summative evaluation scores, proficiency rates will be more sensitive to marginal changes in performance measure weights than in evaluation systems where the proficiency threshold is located at the upper end of the score distribution (where the density of teachers is lower). We conclude by discussing the implications of these findings for policy and practice.

## The Anatomy of a Teacher Evaluation System

The process of assigning a summative evaluation rating to teachers is shaped by four primary design features of a teacher evaluation system: (a) the teacher performance measures used; (b) the approach used to place performance measures on a common scale**;** (c) the weights assigned to teacher performance measures; and (d) the performance ratings thresholds. We describe each of these design features in detail below.

### Teacher Performance Measures

A key feature of newly implemented evaluation systems is the incorporation of multiple measures of teacher performance. In this paper, we focus on three distinct and widely used measures: observations of a teacher's classroom instruction; a teacher's contribution to student achievement growth; and students' perceptions of teacher

effectiveness. Based on a recent analysis documenting the extent of teacher evaluation reform, all 46 states and 23 districts (of the largest 25 school districts and DC) that have, or plan to have, implemented new teacher evaluation systems no later than the 2016–2017 school year incorporate classroom observation as a measure of teacher performance. Further, 80% of these states and districts use one or more measures of teacher performance based on student achievement.[3] Finally, 17% of these states (8) and districts (4) incorporate student surveys capturing students' perceptions of teacher performance (Steinberg & Donaldson, 2016).

**Classroom observations.** Observation rubrics provide scales for criterion-based assessments of a teacher's classroom instruction and professional practice. Evaluation system designers first select among classroom observation protocols (FFT, CLASS, PLATO, etc.) and decide whether to incorporate the full protocol (i.e., all observation components across multiple domains of teacher practice) or a subset of the domains. Next, designers decide on the number of formal/informal classroom observations that each teacher is subject to, and who (e.g., principal, assistant principal, master teacher) is responsible for conducting the classroom observation and rating a teacher's instructional and professional practices on the chosen observation rubric. Evaluation scores from multiple observations are then combined to construct a final teacher practice score.[4]

**Contributions to student achievement.** Measures of teacher performance based on student achievement rely on student test scores and aim to capture measures of student growth attributable to the teacher's instructional performance. Evaluation system designers choose a particular statistical approach for calculating a teacher's contribution

to student learning based on state-administered standardized exams; such approaches include teacher-level VAM and/or student growth percentiles (SGP).[5] These norm-referenced measures capture teachers' contributions relative to their peers, rather than on an absolute scale. Since upwards of 70% of teachers nationwide do not teach in grades and/or subjects in which state-administered exams are available (Watson, Kraemer, & Thorn, 2009), many systems also incorporate criterion-based student learning objectives (SLOs) to evaluate a teacher's contribution to student learning.

**Student surveys.** Student feedback on teacher performance is captured by student perspective surveys. These surveys ask students to report about their teacher's performance and objective occurrences of specific instructional practices. Designers select among a variety of surveys (such as the Tripod survey), and then determine how to construct scores based on students' responses to create these criterion-based measures.

**Placing Teacher Performance Measures on a Common Scale**

Once a teacher has been evaluated on multiple performance measures, consideration must be given to how to place these different measures, which typically vary in how they are scored, onto the same scale. For example, classroom observations that use the FFT observation rubric are scored on an integer scale from 1 to 4 (i.e., a range of 3). In contrast, VAM scores have no theoretical minimum or maximum value (i.e., an infinite range), and the mean score is centered at zero. By placing teacher performance measures on a common scale, weights can be applied to each performance measure to construct a teacher's summative evaluation score. Then, ratings thresholds

can be applied to the summative evaluation score to determine a teacher's summative evaluation rating.

District policymakers therefore determine (a) the point range for the common scale and (b) the mapping of points from different measures onto a common scale. In practice, there exists considerable variation in the point range assigned to a teacher's summative evaluation score. Table 1 provides the range of available evaluation points across a purposeful sample of eight large and geographically diverse districts that have newly implemented teacher evaluation systems. For example, available evaluation points in Chicago Public Schools range from 100-400; in New York City, available evaluation points range from 0-100; in Philadelphia, available evaluation points range from 0-3. Importantly, the distribution of teacher proficiency – which will depend on the performance measure weights and ratings thresholds – will be invariant to the choice of the range of a common point scale. District policymakers typically assign points to each performance measure on a one-to-one basis, since the weight applied to different performance measures allows for local preferences to guide decisions about which measure should have more (or less) influence on a teacher's summative evaluation rating.

**Performance Measure Weights**

After multiple performance measures have been selected and scores have been placed on a common scale, designers must decide how to combine scores into a single summative evaluation score. In the vast majority of systems, this is done by assigning weights (relative proportions of a teacher's summative evaluation score) to each performance measure. For example, if we were to randomly select a teacher teaching in a

tested grade/subject across the nation's largest school districts with newly implemented evaluation systems (i.e., the typical teacher in a tested grade/subject nationwide), 82% of his/her evaluation score (and subsequent summative evaluation rating) will be based on the three performance measures we use in our analyses: classroom observations of teacher practice (52%), student performance on state-administered exams (28%), and student surveys (2%). The balance of this teacher's evaluation will depend on other measures of teacher performance, including SLOs, schoolwide achievement, professional conduct and/or parent/caregiver surveys (Steinberg & Donaldson, 2016). In some evaluation systems, scores are not aggregated into a single summative evaluation score, but instead are mapped from multiple performance measures onto a rating category using a ratings matrix (e.g., Gwinnett County Public Schools in Table 1).

**Performance Rating Thresholds**

Given a teacher's summative evaluation score, a teacher's performance rating in a given school year is most often determined by an evaluation system's ratings thresholds.[6] These thresholds are based on the percent of available evaluation system points that a teacher earns for his or her performance across multiple teacher performance measures. The percent of available evaluation system points that a teacher earns may be calculated as:

(1) $$\textit{Percent of Available Points Earned} = \frac{(Summative\ Evaluation\ Score - Minimum\ Score)}{(Maximum\ Score - Minimum\ Score)}$$

For example, if the evaluation system point scale ranges from a minimum of 1 to a maximum of 4 points (i.e., a scale range of 3) and a teacher's summative evaluation score is 2.5, then a teacher has earned 50% of available evaluation system points [ i.e.

(2.5-1)/(4-1)]. An important implication of this evaluation system design feature is that, once a teacher has been evaluated and has earned his/her evaluation points (and, by extension, the percent of available points in a given evaluation system), the same teacher may be rated differently depending on where the system sets its ratings thresholds.

As shown in Table 1 (and accompanying Figure 1), newly implemented teacher evaluation systems assign quite different thresholds to determine teachers' summative ratings. For example, a teacher who earns 60% of available (district-specific) evaluation points on his/her summative evaluation score would be rated the lowest possible rating level based on New York City's evaluation system, the second lowest rating level in Chicago, Denver, and Miami Dade, but proficient (Level 3) in Clark County, Fairfax County, Gwinnett County, and Philadelphia. A teacher in New York City must earn at least 74% of available evaluation points to be rated proficient/effective (Level 3 in each district), while a teacher in Philadelphia must earn 50% of available evaluation points to receive the same rating. Such variation suggests that districts may differ in both their views concerning what it means for teachers to meet proficiency standards as well as the degree of difficulty in earning evaluation score points across different evaluation systems. Districts, may, for example, adjust their ratings thresholds to correspond to the degree of difficulty of earning points on the district-specific evaluation measures. This may result in districts with vastly different ratings thresholds having quite similar distributions of teacher performance. In our simulation-based analyses described below, we hold constant the set of performance measures used to better illustrate how different ratings thresholds shape teacher proficiency rates.

<Figure 1 about here>

<Table 1 about here>

## Data and Sample

We use data from the Measures of Effective Teaching (MET) study, which was carried out over 2 school years (2009–2010 and 2010–2011) and across six districts.[7] The MET study is among the most ambitious efforts to date to systematically measure teacher effectiveness, and affords a unique opportunity to examine the sensitivity of teacher performance ratings. In particular, the MET data contain a wide range of performance measures that are common to more than 1,000 teachers and which are currently being incorporated into new teacher evaluations systems, including measures based on teacher practice, student achievement, and student reports. Teacher practice is measured using multiple classroom observation protocols; we use scores from Danielson's Framework for Teaching (FFT) protocol given its widespread adoption across newly implemented teacher evaluation systems (Garrett & Steinberg, 2015; Steinberg & Donaldson, 2016). Teacher performance based on student achievement is measured by VAM scores calculated by MET Project researchers. Teacher performance based on students' reports is measured using student responses on the Tripod survey. In the next section we discuss how we construct scores for each teacher from the FFT, VAM, and Tripod survey data.

The teacher sample includes 1,275 teachers in grades 4–8 who participated in the MET study during the 2009–2010 school year. We focus on the first year of the MET study because all teachers were assigned to classes by the normal, within-school process, in contrast to the second year, when many MET teachers were randomly assigned to

classes just prior to the start of the 2010–2011 school year. We are interested in how teacher ratings may be sensitive to weighting schema under conditions in which teachers are assigned to their classes in the typical manner (that is, nonrandomly). Table 2 summarizes the characteristics of teachers included in the sample. All simulations are based on the full sample of teachers (n=1,275).

<Table 2 about here>

## Empirical Approach

### Constructing Performance Measure Scores

We begin by constructing a score for each teacher on each of the three performance measures—teacher practice, teacher contributions to student achievement, and students' reports of teacher practice—as described below.

**Classroom observations of instructional practice.** The MET project used an abbreviated version of The Danielson Framework for Teaching (FFT) observation protocol, including eight components across two domains – Domain 2 (the classroom environment) and Domain 3 (instruction) – with each component rated on a 1 (unsatisfactory) to 4 (distinguished) integer scale. Scores for each of the eight FFT components were generated by MET raters from videos of subject-specific (e.g., math or ELA) lessons that MET teachers conducted on multiple occasions during the 2009–2010 school year.[8] We average across FFT components within lesson observations and then average across lesson observations (within a teacher) to generate a teacher's practice score. We create both a subject-specific practice score ($FFT_{is}$ ) for teacher *i* observed teaching lesson subject *s* (math or ELA) as well as an aggregate practice score

($FFT_{i,Aggregate}$) across all lessons and subjects.[9] This simple approach is used by most school districts to construct classroom observation scores, and prior research using the MET data has found this to be an appropriate approach for aggregating teacher effectiveness measures based on classroom observation scores (Garrett & Steinberg, 2015; Kane et al., 2013; Mihaly, McCaffrey, Staiger, & Lockwood, 2013).[10]

We find that MET teachers received an average FFT score of 2.5 (see Table 3), approximately half a point (and more than one standard deviation) lower than mean FFT scores received in newly implemented teacher evaluation systems (e.g., Chicago Public Schools [Jiang & Sporte, 2016; Steinberg & Jiang, 2016] and Pennsylvania [Lipscomb, Terziev, & Chaplin, 2015]). This is likely the result of several factors. First, MET raters were not physically present in the classroom, as is the case with school-based evaluators. Second, MET raters had no personal connections to the teachers they rated remotely, and did not participate in either pre- or post-observation meetings with teachers as is the practice in many newly implemented evaluation systems. Third, MET ratings of teacher practice were not tied to consequential, high-stakes teacher personnel decisions. Research has found that evaluators systematically assign higher summative ratings to teachers relative to formative ratings which are decoupled from high-stakes consequences (Kraft & Gilmour, in press). Fourth, MET raters received extensive training and were required to pass certification tests in order to conduct remote observations. Finally, MET raters used an abbreviated version of the FFT instrument, which did not require them to evaluate teachers on domains such as Planning and Preparation or Professional Responsibilities. Recent evidence from Baltimore Public Schools indicates that school-

based evaluators (i.e., principals and assistant principals) rate teacher practice, based on classroom observations scores, higher than evaluators who are external to the teacher's school (Jackson & Steinberg, 2017).

To more closely approximate the consequential ratings teachers receive in the context of newly implemented evaluation systems, we adjust MET teachers' FFT scores by adding 0.5 points (which we refer to as $FFT^{A}$). By shifting the mean of the FFT scores, we are able to better approximate (though not replicate) the distribution of teachers' summative evaluation ratings in newly implemented teacher evaluation systems and to more clearly illustrate how performance measure weights and ratings thresholds shape the distribution of teacher effectiveness. We do not adjust the variance of the underlying FFT scores given by external MET raters since we find that the variance in MET FFT scores is no greater (and in some cases lower) than the variance of observation scores found in newly implemented systems in Chicago (Jiang & Sporte, 2016; Steinberg & Jiang, 2016) and Pennsylvania (Lipscomb et al., 2015). By not upwardly adjusting the variance of MET FFT scores, we avoid overstating the effective weight – which is increasing in the variance of the underlying teacher performance measure – given to observation scores by external raters in the MET data (Schochet, 2008). Our substantive findings presented below are not sensitive to adjusting the mean of the FFT scores.

<Table 3 about here>

**Student reports of teacher practice.** Students' reports of their teachers' practices were captured using the Tripod Elementary and Secondary surveys developed by Ron Ferguson. Both versions of the Tripod survey are organized around seven

domains—the 7Cs—of a teacher's classroom practice (i.e., care, control, clarify, challenge, captivate, confer, and consolidate). Students respond to 36 items on a 5-point Likert scale ranging from No, Never to Yes, Always (Elementary) or Totally True to Totally Untrue (Secondary). Following the practices of the Tripod project and the MET Project, we constructed an overall measure of students' assessments of their teachers' instructional practices by assigning point values of 1 to 5 to Likert scale responses, reverse coding items with negative valence, averaging responses across the 36 items for each student, and averaging students' overall scores to the teacher level ($Survey_i$) (For further details see Kane & Cantrell, 2010).[11]

**Teacher contributions to student achievement.** VAM scores were created for the MET sample of teachers using student achievement data from state-administered accountability exams. MET researchers estimated VAMs by grade and district for a single achievement outcome (ELA or math). Student achievement was modeled as a function of student background characteristics and prior-year achievement, in addition to average class background characteristics and prior-year achievement. Residuals from these models were then averaged to generate subject-specific teacher VAM scores ($VAM_{is}$).[12] For subject-matter specialists teaching more than one section of the same subject, we created a weighted average VAM score, weighted by the number of students tested in each of the teacher's sections.

**Placing Measures on a Common Scale**

We rescaled teachers' VAM and Survey scores so that they share the same theoretical and continuous four point scale (i.e., 1 to 4) as the FFT (with corresponding

three-point range). As described above, this is a necessary step for applying weights, but the choice of a common scale does not affect the distribution of teacher proficiency.

A simple linear transformation allows us to rescale the Survey measure as follows:

$$(2)\ Survey_i^{Rescale} = \left( Survey_i * \left( \frac{3}{Survey^{TheoreticalRange}} \right) \right) + \left( 1 - \left( Survey^{TheoreticalMin} * \left( \frac{3}{Survey^{TheoreticalRange}} \right) \right) \right)$$

In Equation (2), $Survey_i$ is the overall score from student surveys for teacher *i* based on students' responses to the 34-item Tripod survey. The value of $Survey^{TheoreticalRange}$ equals 4 (between 1 and 5) and $Survey^{TheoreticalMin}$ equals 1, reflecting the minimum value on the 5-point Likert response scale. The first term on the right side of the equation rescales the range of all teacher Survey scores to equal three, and the second term shifts the score range upward so that the minimum value of the rescaled Survey score ($Survey_i^{Rescale}$) equals 1. This approach preserves the relative position of teachers' empirical Survey scores within the full theoretically possible range.

The approach taken in Equation (2) is not possible for teachers' VAM scores because VAM is a relative measure with no true theoretical range or minimum value. Thus, we substitute empirical analogues into Equation (2) as follows:

$$(3)\ VAM_{is}^{Rescale} = \left( VAM_{is} * \left( \frac{3}{VAM_s^{ObservedRange}} \right) \right) +$$

$$+\left(1-\left(VAM_s^{ObservedMin}*\left(\frac{3}{VAM_s^{ObservedRange}}\right)\right)\right)$$

In Equation (3), $VAM_{is}$ is teacher *i*'s VAM score in subject *s* (math or ELA). The variable $VAM_s^{ObservedRange}$ is the observed range of VAM scores in subject *s* among all teachers in the sample. The term $VAM_s^{ObservedMin}$ is the observed minimum value of VAM scores in subject *s* among all teachers in the sample. As in Equation (1), the first term on the right side of the equation rescales the range of all teacher VAM scores for subject *s* to equal 3, and the second term shifts the score range upwards so that the minimum value of the rescaled VAM score for subject *s* ($VAM_{is}^{Rescale}$) equals 1. We rescale math and ELA VAM scores separately, and, for generalist teachers, we create an aggregate VAM score ($VAM_{i,Aggregate}^{Rescale}$) by averaging teacher *i*'s rescaled VAM math ($VAM_{i,math}^{Rescale}$) and ELA ($VAM_{i,ELA}^{Rescale}$) scores.[13]

Figure 2 shows the score distribution of the three teacher performance measures. Teacher performance was judged to be better by students (based on survey reports) than by external evaluators' observations of teacher practice or student achievement (see Table 3 and Figure 2, Panel A). Interestingly, the distribution of teacher performance based on unadjusted classroom observation scores (*FFT*) is similar to teacher performance based on student achievement measures (*VAM*). However, after shifting the distribution of observation scores upward ($FFT^A$) to more closely reflect how teachers may be rated in the context of newly implemented evaluation systems, we find that the

distribution of observation scores is nearly identical to that of scores based on student survey responses (see Table 3 and Figure 2, Panel B).

<Figure 2 about here>

**Assigning Weights to Performance Measure Scores**

We construct a summative evaluation score for each teacher *i* as a weighted average of the performance measure scores, as follows:

$$(4)\ Score_i^j = (FFT_{i,Aggregate}^{A} * W_{FFT}^{j}) + (VAM_{i,Aggregate}^{Rescale} * W_{VAM}^{j}) + (Survey_i^{Rescale} * W_{Survey}^{j})$$

where $Score_i^j$ is teacher *i*'s summative evaluation score based on weighting scheme *j*, *FFT* is teacher *i's* practice score based on classroom observations, *VAM* is the score that captures teacher *i*'s contribution to student achievement, and *Survey* is teacher *i*'s score based on students' reports of teacher practice. Each of the three performance scores has an associated nominal weight (*W*) that corresponds to weighting scheme *j*. We use the adjusted FFT score ($FFT^A$) in the calculation of all summative evaluation scores.

In practice, teacher evaluation systems may assign any feasible set of nominal (i.e., policy) weights to each of the three performance measure scores, so long as they sum to 100%. The assignment of nominal weights to performance measures has been shown (using MET data) to yield statistically more reliable summative evaluation scores than empirically determined weights (e.g., optimal prediction weights, which are used to predict student test scores); as a result, nominal weights are both better suited for high-stakes evaluation systems and better reflect the relative value that policymakers place on different measures of teacher performance (Martinez et al., 2016). We assign nominal

weights to each of the three performance measure scores based on the following approach. First, we allow $W_{FFT}^{j}$ to vary from 0% to 100% along an integer scale, such that: $W_{FFT}^{j}$ = [0,100]. Next, we construct the weight associated with a teacher's contribution to student achievement as follows: $W_{VAM}^{j} = \left(\frac{100-W_{FFT}^{j}}{(1+Ratio_{Survey/VAM})}\right)$, where $Ratio_{Survey/VAM} = \left(\frac{W_{Survey}^{j}}{W_{VAM}^{j}}\right)$, or the ratio of the weights assigned to the student survey and VAM scores for weighting scheme *j*. The *Ratio* allows for variation in the value that an evaluation system places on students' classroom experiences relative to student achievement as measures of teacher performance. From this, we construct the weight associated with student survey scores as follows: $W_{Survey}^{j} = (100 - (W_{FFT}^{j} + W_{VAM}^{j}))$.

Following this approach, we generate four summative performance scores for each teacher. For the first performance score ($Score_{i,1}^{j}$), we set *Ratio*= $\frac{1}{10}$. This value of $Ratio_{Survey/VAM}$ is motivated by the fact that, among the largest school districts with newly implemented teacher evaluation systems, the average weight assigned to VAM is approximately 10 times the average weight assigned to student survey scores for teachers teaching in tested grades/subjects (Steinberg & Donaldson, 2016).[14] For example, if the entirety of a teacher's summative evaluation score depends on classroom observations (i.e., $W_{FFT}^{j}$=100), then zero weight will be assigned to the VAM and student survey scores. If, however, none of a teacher's summative evaluation score depends on classroom observations (i.e., $W_{FFT}^{j}$=0), and the evaluation system assigns 10 times as

much weight to VAM as it does to the student survey measure, then: $W_{VAM}^{j} = \frac{100 - W_{FFT}^{j}}{(1 + Ratio)} = \frac{100 - 0}{(1 + \frac{1}{10})}$ = 90.9%, and $W_{Survey}^{j}$ = 100 – ($W_{FFT}^{j}$ + $W_{VAM}^{j}$) = 100 – (0 + 90.9) = 9.1%.

To allow for student surveys (and, by extension, students' reports of teacher performance) to play a more prominent role in teachers' summative evaluation scores (relative to student achievement), we construct a second performance score ($Score_{i,2}^{j}$) by setting *Ratio*= $\frac{1}{5}$. This value of $Ratio_{Survey/VAM}$ is motivated by evidence that the weight assigned to VAM is approximately five times the weight assigned to student survey scores, on average, across newly implemented systems (in the largest school districts) that give non-zero weight to student surveys and VAM (Steinberg & Donaldson, 2016). Further, in some evaluation systems, student surveys contribute even more to a teacher's summative evaluation score. Indeed, in some districts, student surveys are assigned approximately half the weight that is assigned to teacher performance based on student achievement.[15] We therefore construct a third performance score ($Score_{i,3}^{j}$) by setting *Ratio*= $\frac{1}{2}$. Finally, many new evaluation systems do not incorporate student surveys into teachers' summative evaluation scores.[16] We construct a fourth performance score that is composed of only observation and VAM scores($Score_{i,4}^{j}$) by setting *Ratio*=0. The incorporation of multiple ratios for teacher performance measures into our analysis allows for greater insight into how the distribution of teacher ratings responds dynamically to the interaction between score construction and the two key system design features – performance measure weights and ratings thresholds.

A performance measure's contribution to a teacher's summative score depends not only on the weight system designers assign to the measure (i.e., nominal weight), but also on the underlying variance of the measure and its correlation with the other measures used to construct the summative score (i.e., effective weight) (Schochet, 2008). As Schochet (2008) notes, equal weight assigned to performance measures does not imply that each performance measure will contribute equally to the overall variance of a teacher's summative score. Specifically, the effective weight for each performance measure will depend on its average correlation with the other performance measures; if the average correlations are similar across measures, then the effective and nominal weights should also be similar (Schochet, 2008). Simply put, measures with lower correlations with other performance measures will have lower effective weights than their nominal (i.e. assigned) weights suggest.

In our analytic sample of MET teachers, we observe that VAM is relatively weakly correlated with both the FFT score (.11) and the student survey score (.17). In contrast, the FFT score is more highly correlated with the student survey score (.41) (see Table 4). For example, suppose equal weights are assigned to each of the three performance measures (i.e., 33.3% assigned to observation scores, VAM and student survey scores). Based on the performance measures in the MET data used in this paper, the effective weights will be as follows: 34.8% to observation scores, 29.2% to VAM, and 36.0% to student surveys (Schochet, 2008).[17] This example illustrates that VAM, which has the lowest correlation with the other performance measures, will also have the lowest effective weight. In practical terms, the measure with the lowest effective weight

will contribute the least to a teacher's summative evaluation score when equal nominal weight is assigned to each teacher performance measure.

<Table 4 about here>

**Examining the Sensitivity of Ratings to System Design**

To examine the sensitivity of teachers' evaluation ratings to evaluation system design parameters, we conduct two sets of simulation-based analyses. For the first analysis, we examine how, under a fixed evaluation ratings system (i.e., ratings thresholds employed in one of the eight teacher evaluation systems), the distribution of teacher ratings may be sensitive to the underlying weights assigned to performance measures. Based on a given ratings system, we vary the weights assigned to the three performance measures and calculate the proportion of teachers who would be rated proficient under each weighting scheme. Teachers are deemed proficient if the evaluation points that they earn are sufficient for them to receive one of the two highest ratings– level 3 or level 4 – which are based on the fixed ratings thresholds of each district's evaluation system. Teachers who achieve at least a level 3 summative evaluation rating are deemed proficient in each of the eight districts included in our analysis (see Table 1).

For the second analysis, we examine how, under a fixed weighting scheme, the distribution of teacher ratings may be sensitive to different ratings threshold schema found across our sample of eight district evaluation systems. We do so by calculating the proportion of teachers who would be rated proficient when only the ratings thresholds vary. This analysis allows us to demonstrate the extent to which teachers who receive the same summative evaluation score (and, by extension, the same percent of total evaluation

points available) may be rated differently as a consequence of policy-determined ratings thresholds. These complementary analyses also allow us to examine how the properties of teacher evaluation measures influence teacher proficiency rates under different system design parameters.

## Results

Table 5 summarizes our primary results. These simulated findings do not (nor are they intended to) replicate the actual ratings distributions in the eight districts from which we draw our ratings thresholds. Indeed, the simulated proficiency rates reported in Table 5 are substantially lower for some districts than the actual proficiency rates found in new evaluation systems (Anderson, 2013; Kraft & Gilmour, in press). This is likely due to a number of factors, including: the specific performance measures used by each district; the norms across districts about what constitutes proficient practice; the exclusion of other types of measures and observation domains; and the consequences and rewards attached to teacher ratings.

To examine how variation in teacher performance measure weights shape the distribution of teacher proficiency, we look within a given teacher evaluation system, allowing us to hold constant the performance ratings thresholds while varying the performance measure weights. First, we find that teacher proficiency rates change substantially as the weights assigned to teacher performance measures change. Looking down a given column, or evaluation system (within a panel of Table 5), we see how the proportion of proficient teachers differs under different component weight schemes. Take the rates of teacher proficiency based on the ratings thresholds of Fairfax County Public

Schools' evaluation system (see Table 5, Panel A, which is based on $Score_1$). Under a component weight scheme where $FFT^A$ receives zero weight (VAM contributes 90.9% and student survey contributes 9.1% to a teacher's summative evaluation score), 45% of teachers in our sample would be rated proficient. If we change only the component weights—say, to 50% $FFT^A$ (and 45.5% VAM and 4.5% student survey)—then teacher proficiency increases to 85%, an increase of 40 percentage points.[18]

<Table 5 about here>

Our findings in Table 5 reveal two important facts with respect to the weight assigned to performance measures with higher mean values. First, the more weight assigned to measures with higher relative means, the greater the rate of teacher proficiency. This can be seen by looking within a given evaluation system (i.e., within a column of Table 5) as the weight for observations scores (FFT) increases relative to VAM scores across all four Survey/VAM ratios (i.e., within a panel of Table 5). Second, when greater relative weight is assigned to measures with lower means – for example, by reducing the Survey/VAM ratio from ½ (in Panel C) to 0 (in Panel D) – assigning more weight to a third measure (FFT) with a higher mean value will produce larger incremental changes in teacher proficiency rates. Specifically, focusing on teacher proficiency rates based on Chicago's system: when the Survey/VAM ratio is the greatest (at ½; see Panel C), increasing the weight assigned to FFT from 50 to 100% increases teacher proficiency rates by 23 percentage points, from 53 to 76%. In contrast, when the Survey/VAM ratio is the lowest (at 0, see Panel D), increasing the weight assigned to FFT from 50 to 100% has a much larger effect on the change in proficiency rates, increasing teacher proficiency

this time by 45 percentage points, from 31 to 76%. This empirical fact bears out across each of the other seven systems with different ratings thresholds.

We further illustrate these results with a series of heat maps in Figure 3. For each of the eight evaluation systems, these figures illustrate how the distribution of teacher ratings changes as the weight assigned to the adjusted observation score ($FFT^A$) increases from 0% to 100%. These figures clearly show how, in a single evaluation system with fixed ratings thresholds, the percentage of teachers assigned to each rating category substantively changes across different weighting schemes.

Figure 3 also demonstrate how performance weights and ratings thresholds interact differently across evaluation systems to determine the distribution of teacher effectiveness (i.e., the proportion of teachers in each performance rating category). Evidence from Figure 3 reveals how changes to the distribution of teacher ratings depend on the specific rating threshold system with which a given set of performance measure weights is combined. Specifically, based on Miami's and NYC's evaluation systems, teacher proficiency rates among our sample of teachers remain relatively constant until $FFT^A$ contributes (approximately) at least 70% of the weight to a teacher's summative evaluation score, after which teacher proficiency increases at a relatively constant rate (see Panels F and G, Figure 3). Based on Denver's evaluation system, teacher proficiency rates remain relatively constant until $FFT^A$ contributes (approximately) at least 20% of the weight to a teacher's summative evaluation score (see Panel C, Figure 3). In contrast, teacher proficiency rates based on the ratings thresholds in evaluation systems located in Chicago, Clark County, Fairfax County, Gwinnett County, and Philadelphia, increase at a

relatively constant rate as the FFT weight increases across the full range of the FFT weight distribution.

<Figure 3 about here>

Second, we find that teacher proficiency rates change substantially when, holding constant the performance measure weights, the same teachers are evaluated using different performance ratings thresholds. By looking across a given row, or weight scheme (within a panel of Table 5), we see how the proportion of proficient teachers differs across evaluation systems. Based on a weight scheme where $FFT^{A}$ contributes 50% to a teacher's summative evaluation score, our simulated teacher proficiency rates range from 3% and 4%—based on the ratings thresholds in Miami's and NYC's evaluation systems, respectively—to approximately 90%— based on the ratings thresholds in Fairfax County's and Philadelphia's systems (see Panel B, Table 5). Figure 4 illustrates these results graphically by capturing the full distribution of teacher ratings across the eight evaluation systems (and across the four score constructions) under a weight scheme where FFT contributes 50% to a teacher's summative evaluation score. Here we see that the proportion of teachers rated in all four categories, and in particular the lowest two rating categories (i.e., levels 1 and 2), vary substantially due to differences across districts' ratings thresholds.

<Figure 4 about here>

Third, we show that the relative weights teacher evaluation systems place on student information – student survey responses relative to student achievement exams – in the construction of a teacher's summative performance score will have real

consequences for the distribution of teacher proficiency rates. Table 6 summarizes the range of teacher proficiency rates, within and across teacher evaluation systems, for different constructions of a teacher's performance score (fixing the weight assigned to observation scores at 50%). For example, we find that lowering the Survey/VAM ratio from ½ to 0 reduces teacher proficiency rates by up to 22 percentage points – from 53 to 31% (based on the ratings thresholds in Chicago's system) and from 61 to 39% (based on the ratings thresholds in Clark County's system). These findings further reveal that teacher proficiency rates are lowest across all systems when norm-referenced teacher performance measures such as VAM are given greater relative weight than criterion-based measures such as student surveys. This result is not surprising given that teachers' VAM scores are, on average, lower than teacher scores based on student surveys (see Table 3 and Figure 2) and have lower correlations with the other teacher performance measures in the MET data (see Table 4).

<Table 6 about here>

## Discussion

Recent policy reforms have spurred a major overhaul of teacher evaluation systems in the United States, highlighted by the incorporation of multiple measures of teacher performance and the expansion of teacher rating categories in an effort to better measure and differentiate teacher effectiveness. The designs of these new systems also incorporate equally important choices that policymakers have made concerning the weights assigned to multiple performance measures and the placement of teachers' summative scores into discrete performance categories. Yet, little guidance has been

available to inform policymakers about the consequences these design decisions may have on the distribution of teacher ratings and the proportion of teachers deemed proficient. The absence of empirically-based guidance to inform these decisions is particularly notable given that teachers' summative ratings are increasingly being used to make high-stakes personnel decisions.

Not only do we find that both the weighting schemes assigned to performance measures and the ratings thresholds set by evaluation systems play a critical role in determining teacher proficiency rates, but also that the properties of performance measures directly influence the distribution of teacher proficiency rates. First, if teacher performance measures are assigned the same weight, then measures that are more highly correlated will effectively contribute more to a teacher's summative rating than measures that are weakly correlated. Second, teacher proficiency rates are increasing in the weight assigned to teacher performance measures with highest mean values.

Further, we show that teacher proficiency rates depend on the relative value a system places on student information – student surveys relative to student achievement – in the construction of a teacher's summative performance score. We also demonstrate that teacher proficiency rates are much more sensitive to changing teacher performance weights when the proficiency ratings threshold is located near the center of the distribution of actual ratings (e.g., Chicago and Clark County) than in systems where the proficiency threshold is located at the upper end of the score distribution (e.g., New York City and Miami).

These results provide new evidence to inform policymakers about how design decisions related to teacher performance measure weights and ratings thresholds affect the distribution of teacher proficiency rates. Our analysis also provides empirical insight into how the properties of teacher evaluation measures shape the distribution of teacher proficiency rates. In doing so, our results point to the fact that variation in teacher proficiency rates can be predictable and quantifiable based on both the observed properties of teacher performance measures and system design decisions related to performance measure weights and ratings thresholds.

## Implications for Policy

Our analyses illustrate several important findings that are particularly salient for policymakers. First, differences in the relative weights assigned to norm-referenced versus criterion-based measures of teacher performance can result in substantially different distributions of teacher performance ratings. Norm-referenced measures, such as VAM, are relative scores that are normalized within a given group of teachers; by construction, the mean of VAM scores will be centered at the middle of the score distribution. In contrast, the mean of criterion-based measures, such as observation and survey scores, can be located anywhere along the score distribution. In practice, criterion-based measures used in teacher evaluation systems are often centered well above the middle of the score range (see Figure 2, Panel B). Therefore, if scores on criterion-based measures are systematically skewed toward the upper end of the score distribution, then the relative weights assigned to norm-referenced versus criterion-based performance measures will have real consequences for the distribution of teachers' summative

evaluation ratings. Indeed, we show that teacher proficiency rates reach a minimum when VAM scores receive the greatest nominal weight (see Tables 5 and 6). Moreover, we also show that performance measures with lower average correlations with the other performance measures, such as VAM scores, will contribute less to teachers' summative evaluation scores – they will have lower effective weights – than its nominal weight would suggest.

These findings suggest that summative evaluation ratings for the nearly 70% of teachers in non-tested grades/subjects (Watson et al., 2009) are likely to differ in systematic ways from the ratings of teachers for whom VAM scores can be calculated. We remind readers that VAM scores were available for all MET study teachers included in our simulation-based analyses. In the MET data, we observe that VAM has a substantially lower mean and greater variance than both observation scores (based on the adjusted FFT ratings of teacher performance) and student survey scores (see Table 3). Further evidence of this pattern is found, for example, in Tennessee's teacher evaluation system (Tennessee Department of Education, 2015). Since greater weight is consistently assigned to observation scores for teachers in non-tested grades and subjects, we would expect these teachers' evaluation ratings to be systematically higher than those for teachers with VAM scores. Data provided to us from a large (anonymous) urban school district in the Midwest bore out this prediction. In the absence of student achievement on state accountability exams, replacing VAMs with measures based on locally developed tests of student progress (i.e., SLOs) is unlikely to resolve this ratings disparity. SLOs are criterion-based measures of teacher performance that are often systematically and

upwardly skewed relative to VAM scores. This also helps to further explain why our simulated teacher proficiency rates understate the share of teachers who, in a typical school district, are deemed to be proficient.

Second, states' and districts' decisions about where to place the summative ratings thresholds have direct implications for the distribution of teacher effectiveness. In systems that apply absolute thresholds—as do the eight evaluation systems included in our analysis—all teachers, in principle, may be rated proficient. In contrast, a system that imposes a target distribution of teacher effectiveness, such as the one used by the Dallas Independent School District, defines the proportion of summative evaluation scores (and, by extension, the proportion of teachers) assigned to a summative evaluation rating.[19] Despite the important differences in teacher performance measures used across districts and the weights assigned to these measures, there is no clear justification for the variation in ratings thresholds that districts set for teachers to meet proficiency standards (as shown in Figure 1). As a result, these differences in ratings thresholds complicate comparisons of teacher effectiveness across districts in the same way that state-specific differences in performance standards limit the comparability of student proficiency rates across states.

Further, both the meaning attached to and the consequences of a teacher's summative rating also differ across districts. For example, a level 2 rating of 'Developing' (as in Chicago, Miami and New York City) may imply that teachers are making progress toward proficiency, while a level 2 rating of 'Minimally Effective' (as in Clark County) may imply that they are not. As a result, the district-specific distribution of teacher proficiency will depend, in part, on how different districts assign different

meaning, and, ultimately, different high-stakes consequences to teacher ratings. Our analysis provides insight into how a district's design decisions may influence the desired distribution of teacher ratings.

Third, the design and consequences of evaluation systems may be shaped in important ways by political considerations as well as local implementation practices. Though our simulation-based results reveal that the same teachers may be assigned to different performance categories depending on district-specific ratings thresholds, teacher proficiency rates do not differ in practice nearly as much. This is because system designers may start by selecting ratings thresholds based on objective criteria but then, upon inspection of the distribution of teacher proficiency they generate, revise the ratings thresholds to produce a distribution of teacher proficiency that is both professionally as well as politically feasible. Further, school-based evaluators can respond to a system established by state and district administrators by adjusting scores on subjective teacher performance measures (such as observation scores) in order to produce a desired ratings distribution. Such policy and practice decisions may be made to avoid negative externalities (e.g., teacher exits among high-performing, lower-rated teachers and lower staff moral) that can result when teacher ratings are not consistent with perceptions of effectiveness among educators, even in systems designed to generate greater variation in teachers' summative ratings than what existed under traditional evaluation systems.

Finally, our work calls for greater attention to and transparency around the policymaking process for selecting performance measure weights and determining ratings thresholds. The empirical findings of this paper reveal the sensitivity of teacher ratings to

these design features of newly implemented evaluation systems. These findings demonstrate why it is important that policymakers both understand the consequences of their design decisions and more clearly communicate how these decisions are made. Such considerations are critical in light of policy efforts to improve teacher quality and, ultimately, student achievement.

## Notes

[1] Under the district's traditional evaluation system, nearly all Chicago Public Schools teachers (93 percent) were rated proficient (i.e., Superior or Excellent) under the district's traditional evaluation system, while only 34 percent of Chicago Public Schools met state proficiency standards (Steinberg & Sartain, 2015).

[2] Specifically, 83 percent of states with newly implemented evaluation systems link teacher summative ratings to required professional development for low-rated teachers; among the largest districts and the District of Columbia (DC), 74 percent require professional development for low-rated teachers. Further, 61 percent of newly implemented state evaluation systems (and 39 percent of newly implemented systems in the largest school districts and DC) tie teacher ratings to employment termination, while 48 percent of new state systems and 22 percent of new district systems tie teacher ratings to tenure granting/revocation decisions (Steinberg & Donaldson, 2016).

[3] Approximately 61% (14) of the largest districts and 30% (14) of states use VAM, while 35% (8) of the largest districts and 59% (27) of states use student growth percentiles (Steinberg & Donaldson, 2016).

[4] In practice, many states and districts construct a final teacher practice score by averaging across protocol components (within a given observation) and then averaging across multiple observations. Indeed, prior research suggests that a simple average is an appropriate approach for aggregating teacher practice scores across multiple classroom observations (Garrett & Steinberg, 2015; Kane et al., 2013; Mihaly, McCaffrey, Staiger, & Lockwood, 2013). Alternatively, some evaluation systems weight observation components differently (e.g., Chicago Public Schools' *REACH* system gives greater weight to observation components related to aspects of a teacher's instructional performance [such as engaging students in learning] than to aspects of a teacher's practice related to managing the classroom environment [such as managing student behavior]).

[5] Though the statistical models used to produce teacher VAM scores differ, in general, such models provide an estimate of a teacher's contribution to student learning by controlling for prior student achievement and other observable student and classroom characteristics. We further note that the necessity of a student's baseline test score to construct VAM often precludes teachers in early elementary grades (i.e., grades K-3) from receiving VAM scores. Student growth percentiles provide an estimate of a teacher's contribution to student learning by comparing student achievement growth to students' peers with similar prior test score histories. SLOs are subject- and grade-specific learning goals, which tend to be based on locally selected (i.e., school and/or district) measures of student achievement, and which are used to estimate a teacher's contribution to student learning in grades/subjects that are not tested via the state's accountability exam (Steinberg & Donaldson, 2016).

[6] Some districts allow evaluators to use their professional judgement to assign a summative rating based on their own synthesis of multiple performance measures available for each teacher (e.g. Boston Public Schools).

[7] The six participating districts were Charlotte-Mecklenburg Schools (NC), Dallas Independent School District (TX), Denver Public Schools (CO), Hillsborough County Public Schools (FL), Memphis City Schools (TN), and the New York City Department of Education (NY).

[8] Classroom generalist teachers were videoed, on average, on 4 separate days throughout the year, with each day producing one ELA and one math lesson video. Subject-specific (i.e., departmentalized) teachers were videoed on 2 separate days, capturing two different sections of the same subject taught by the teacher.

[9] Aggregate practice scores for departmentalized teachers' (i.e., teachers teaching multiple sections of the same subject) will equal their subject-specific practice scores.

[10] We also pursued measurement-based approaches—Empirical Bayes and Principal Components Analysis—as alternative ways of constructing teacher performance scores using classroom observation scores. The scores generated by these measurement-based

approaches are all very highly correlated with scores constructed by averaging across observation components and domains.

[11] We also pursued alternative approaches to constructing teacher performance scores based on students' perceptions. These include (a) averaging within the 7C domains first and then averaging across these domains so that each domain is weighted equally; (b) averaging individual items across students for each teacher and then averaging across items; and (c) a measurement-based approach where items are weighted based on the loadings from the first Eigenvector of a principal components analysis. All three alternative approaches produce performance scores that are correlated at 0.98 or above. Further sensitivity analyses including dropping students who "straight line" (i.e., fill in the same answer for every single item—less than 0.0001 percent of students) or students whose rating of a teacher is beyond two standard deviations (approximately 3% of students) from the class mean does not change a teacher's performance score.

[12] For more details on the construction of teacher VAM scores, see White & Rowan (2012).

[13] For departmentalized teachers (i.e., teachers teaching multiple sections of the same subject), their subject-specific VAM scores equal their aggregate VAM scores. Averaging across ELA and math VAM scores for a generalist teacher is a practice used in many school districts (e.g., Chicago Public Schools).

[14] For the typical teacher teaching in a tested grade/subject, 21.7% and 1.9% of a teacher's summative evaluation score is based on VAM and student survey scores, respectively (Steinberg & Donaldson, 2016).

[15] For example, for most grade 3-12 teachers in the Dallas Independent School District, student surveys (based on students' classroom experiences) and student-achievement based measures account for 15 and 35 percent, respectively, of a teacher's summative evaluation score (source: Teacher Excellence Initiative, http://tei.dallasisd.org/home-2/defining-excellence/).

[16] Some of the largest school districts, including New York City and Chicago Public Schools, do not incorporate student surveys into teachers' summative evaluation scores (see Table 1).

[17] Following Schochet (2008), the effective weight – the contribution of performance measure *j* to the variance of the composite teacher evaluation score – may be calculated as: $w_j^{effective} = (w_j^{nominal})^2 + \sum_{j \neq k}^{N} w_j^{nominal} w_k^{nominal} \rho_{jk}$ , where *nominal* indicates the nominal weight assigned to performance measure *j* and $\rho_{jk}$ is the correlation between performance measure *j* and performance measure *k* (e.g., the correlation between the FFT score and the VAM score).

[18] Under an evaluation system with very different ratings thresholds, changing the underlying component weights may have little consequence for teacher proficiency. That is, as we vary the component weight scheme, teacher proficiency changes differently depending on where a given evaluation system sets its performance ratings thresholds. Contrast the results based on Fairfax County's system with New York City's system (see Figure 1 and Table 1 for the contrast in ratings thresholds across districts). Under a component weight scheme where $FFT^A$ receives zero weight, 2% of teachers in our sample would be rated proficient based on NYC's ratings thresholds. If we change only the component weights to 50% $FFT^A$ (and 45.5% VAM and 4.5% Survey), teacher proficiency among our sample of teachers would increase to just 3%, based on the ratings thresholds under NYC's evaluation system.

[19] For the 2014–2015 and 2015–2016 school years, Dallas Independent School District's (ISD) Teacher Excellence Initiative (TEI) determined teachers' summative evaluation ratings by arranging scores to follow a target distribution, as follows: 3% of teachers were rated Unsatisfactory; 37% Progressing; 58% Proficient; and 2% Exemplary (Dallas Independent School District, 2015; Dallas Independent School District, 2016).

## References


Almy, S. (2011). *Fair to everyone: Building the balanced teacher evaluations that educators and students deserve*. Washington, DC: Education Trust.

Anderson, J. (2013, March 30th). Curious Grade for Teachers: Nearly All Pass. *The New York Times*. Retrieved from: http://www.nytimes.com/2013/03/31/education/curious-grade-for-teachers-nearly-all-pass.html

Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Cullen, J. B., Koedel, C., & Parsons, E. (2016). *The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality* (No. w22805). National Bureau of Economic Research.

Dallas Independent School District. (2015, May 18). Defining excellence: Rules and procedures for calculating achievement statistics, evaluation scores, and effectiveness levels for Dallas ISD's Teacher Excellence Initiative. Retrieved on June 19, 2015 from http://www.dallasisd.org/Page/28269

Dallas Independent School District. (2016, January 20). Defining excellence: Rules and procedures for calculating achievement statistics, evaluation scores, and effectiveness levels for Dallas ISD's Teacher Excellence Initiative. Retrieved on March 10, 2016 from http://tei.dallasisd.org/home-2/resources/

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. New York: Teachers College Press.

Dee, T., & Wyckoff, J. (2015). Incentives, selection and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267-297.

Donaldson, M. L. (2009). *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality*. Center for American Progress. https://cdn.americanprogress.org/wp-content/uploads/issues/2009/06/pdf/teacher_evaluation.pdf

Donaldson, M. L., & Papay, J. P. (2015). Teacher evaluation for accountability and development. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy*. New York: Routledge.

Elmore, R. F. (2002). Unwarranted intrusion. *Education Next*, *2*(1).

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, *37*(2), 224–242.

Grissom, J. A., & Youngs, P. (Eds.). (2015). *Improving teacher evaluation systems: Making the most of multiple measures*. New York: Teachers College Press.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, *116*(1), 1–28.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64.

Jackson, C., & Steinberg, M.P. (2017). Does teacher effectiveness depend on who rates classroom practice? Evidence from an urban teacher preparation program. *Working Paper*.

Jiang, J. Y., & Sporte, S. (2016). *Teacher evaluation in Chicago: Differences in observation and value-added scores by teacher, student, and school characteristics.* Chicago: University of Chicago Consortium on School Research.

Kane, T., & Cantrell, S. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project* (Bill & Melinda Gates Foundation MET Project Research Paper).

Kane, T., Kerr, K., & Pianta, R. (2014). *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project*. New York: John Wiley & Sons.

Kane, T., McCaffrey, D., Miller, T., & Staiger, D. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (Bill & Melinda Gates Foundation MET Project Research Paper).

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains (Bill & Melinda Gates Foundation MET Project Research Paper).

Kraft, M. A., & Gilmour, A. (in press). Revisiting the widget effect: Teacher evaluation reforms and distribution of teacher effectiveness ratings. *Educational Researcher*.

Le Floch, K. C., Boyle, A., & Therriault, S. B. (2008). *Help wanted: State capacity for school improvement* (AIR Research Brief). American Institutes for Research.

Lipscomb, S., Terziev, J., & Chaplin, D. (2015). *Measuring teachers' effectiveness: A report from Phase 3 of Pennsylvania's pilot of the Framework for Teaching.* Mathematica Policy Research.

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement*. ASCD.

Martinez, J.F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis, 38*(4), 738-756.

McGuinn, P. (2012). Stimulating reform: Race to the Top, competitive grants, and the Obama education agenda. *Educational Policy*, *16*(1), 136–159.

Mihaly, K., McCaffrey, D. F., Staiger, D., & Lockwood, J. R. (2013). *A composite estimator of effective teaching* (Technical report for the Measures of Effective Teaching project.) Seattle, WA: Bill & Melinda Gates Foundation.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation: Lessons learned from observations, principal-teacher conferences, and district implementation.* Chicago: Consortium on Chicago School Research.

Sartain, L., & Steinberg, M.P. (2016). Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago public schools. *Journal of Human Resources*, *51*(3), 615-655.

Schochet, P. Z. (2008). *Technical methods report*: *Guidelines for multiple testing in impact evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Sinnema, C., & Robinson, V. (2007): The Leadership of Teaching and Learning: Implications for Teacher Evaluation. *Leadership and Policy in Schools, 6*(4), 319–343.

Steinberg, M. P., & Donaldson, M. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, *11*(3), 340-359.

Steinberg, M.P., & Jiang, J. (2016). Rater bias or teacher sorting? Examining the causes and consequences of racial gaps in teacher performance ratings. *Working Paper*.

Steinberg, M.P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy,* 10(4), 535-572.

Stronge, J. H., & Tucker, P. D. (2003). *Teacher evaluation. Assessing and improving performance.* Larchmont, NY: Eye on Education.

Tennessee Department of Education. (2015). *Teacher and administrator evaluation in Tennessee: A report on Year 3 implementation.* Retrieved from: http://team-tn.org/wp-content/uploads/2013/08/rpt_teacher_evaluation_year_31.pdf

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector.

Tucker, P. D. (1997). Lake Wobegon: Where all teachers are competent (or, have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education, 11*, 103–126.

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod

student perception survey. *American Educational Research Journal*, *53*(6), 1834-1868.

Watson, J. G., Kraemer, S. B., & Thorn, C. A. (2009). *The other 69 percent*. Washington, DC: Center for Educator Compensation Reform, U.S. Department of Education, Office of Elementary and Secondary Education.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn: The New Teacher Project.

White, M., & Rowan, B. (2012). *A user guide to the "core study" data files available to MET early career grantees*. Ann Arbor: Inter-University Consortium for Political and Social Research, The University of Michigan.

**Table 1. Overview of Teacher Evaluation System Designs**

| | | | | Summative Performance Ratings | | |
|---|---|---|---|---|---|---|
| **District (State)** | **Evaluation System (Year of Study)** | **District Ranking (Size)** | **Teacher Performance Measure (Weight)** | **Scale Range (Available Points)** | **Threshold (% of Available Points)** | **Category (Level)** |
| Chicago Public Schools (IL) | Recognizing Educators Advancing Chicago's Students (2014-15) | 3rd | Observation (70%) | 100-400 (300 points) | 0-36% | Unsatisfactory (1) |
| | | | Student Achievement (30%) | | 36-61% | Developing (2) |
| | | | Student Survey (0%) | | 61-80% | Proficient (3) |
| | | | Other (0%) | | 80-100% | Excellent (4) |
| Clark County School District (NV) | Nevada Educator Performance Framework (2014-15) | 5th | Observation (100%) | 1-4 (3 points) | 0-30% | Ineffective (1) |
| | | | Student Achievement (0%) | | 30-60% | Minimally Effective (2) |
| | | | Student Survey (0%) | | 60-86% | Effective (3) |
| | | | Other (0%) | | 86-100% | Highly Effective (4) |
| Denver Public Schools (CO) | Leading Effective Academic Practice (2014-15) | 34th | Observation (40%) | 0-50 (50 points) | 0-47% | Not Meeting (1) |
| | | | Student Achievement (50%) | | 47-64% | Approaching (2) |
| | | | Student Survey (10%) | | 64-81% | Effective (3) |
| | | | Other (0%) | | 81-100% | Distinguished (4) |
| Fairfax County Public Schools (VA) | Teacher Performance Evaluation Program (2012-13) | 11th | Observation (60%) | 10-40 (30 points) | 0-30% | Ineffective (1) |
| | | | Student Achievement (40%) | | 30-50% | Developing/Needs Improvement (2) |
| | | | Student Survey (0%) | | 50-80% | Effective (3) |
| | | | Other (0%) | | 80-100% | Highly Effective (4) |
| Gwinnett County Public Schools (GA) | Gwinnett Teacher Effectiveness System (2014-15) | 13th | Uses matrix rather than weights for Observation and Student Achievement | 0-30 (30 points) | 0-20% | Ineffective (1) |
| | | | | | 20-53% | Needs Development (2) |
| | | | | | 53-87% | Proficient (3) |
| | | | | | 87-100% | Exemplary (4) |

| | | | | | | |
|---|---|---|---|---|---|---|
| Miami-Dade County Public Schools (FL) | Instructional Performance Evaluation and Growth System (2014-15) | 4th | Observation (50%)<br>Student Achievement (35%)<br>Student Survey (0%)<br>Other (15%) | 0-100 (100 points) | 0-36%<br>36-73%<br>73-88%<br>88-100% | Unsatisfactory (1)<br>Developing (2)<br>Effective (3)<br>Highly Effective (4) |
| New York City Department of Education (NY) | NYC Advance (2013-14) | 1st | Observation (60%)<br>Student Achievement (40%)<br>Student Survey (0%)<br>Other (0%) | 0-100 (100 points) | 0-64%<br>64-74%<br>74-90%<br>90-100% | Ineffective (1)<br>Developing (2)<br>Effective (3)<br>Highly Effective (4) |
| School District of Philadelphia (PA) | Educator Effectiveness System (2012-13) | 19th | Observation (50%)<br>Student Achievement (50%)<br>Student Survey (0%)<br>Other (0% ) | 0-3 (3 points) | 0-16%<br>16-50%<br>50-83%<br>83-100% | Failing (1)<br>Needs Improvement (2)<br>Proficient (3)<br>Distinguished (4) |

Notes. See Data Appendix for full details about data sources. District ranking (size) is based on district enrollment (source: National Center on Teacher Quality: http://www.nctq.org/districtPolicy/contractDatabase/customReport.do#criteria). Teacher Performance Measures (and associated weights) are for teachers in tested grades/subjects. For evaluation systems that evaluate teachers on professional responsibility (Clark County (NV), Denver (CO), Fairfax County (VA), Miami (FL)), we have included the weight assigned to professional responsibility as part of the weight assigned to a teacher's observation score. Some systems incorporate multiple measures of student achievement (in addition to value-added measures) into teachers' summative evaluation scores (e.g., Chicago (IL), New York City (NY)); we have aggregated all student achievement-based measures of teacher performance into the weight assigned to Student Achievement. In some evaluation systems, performance measures other than observation, student achievement or student surveys are incorporated into teachers' summative evaluation scores (e.g., professional development plan as in Miami (FL)).

**Table 2. Teacher Characteristics**

| Teacher Characteristic | All Teachers | Math Teachers | ELA Teachers |
|---|---|---|---|
| Female | .83 | .82 | .87 |
| White | .58 | .56 | .58 |
| Black | .35 | .37 | .36 |
| Hispanic | .05 | .05 | .05 |
| Other | .02 | .02 | .01 |
| Grade 4 | .22 | .28 | .28 |
| Grade 5 | .23 | .30 | .31 |
| Grade 6 | .21 | .16 | .16 |
| Grade 7 | .17 | .13 | .12 |
| Grade 8 | .17 | .13 | .13 |
| Experience (total) | 10.7<br>(8.89) | 10.9<br>(9.56) | 10.4<br>(8.43) |
| Experience (district) | 7.6<br>(6.85) | 7.2<br>(6.63) | 7.3<br>(6.62) |
| Masters+ | .36 | .42 | .40 |
| Generalist | .34 | .52 | .49 |
| Teachers | 1275 | 833 | 874 |
| Schools | 207 | 189 | 196 |
| Districts | 6 | 6 | 6 |

Notes. Proportions reported for all characteristics except Experience, which reports mean (standard deviation). Data are from the 2009–2010 school year. Generalist teachers are included in both the Math and ELA teacher samples. For the full teacher sample, 1,237 teachers reported gender, 1,235 reported race, 580 reported years of experience (total), 992 reported years of experience (in district), and 993 reported educational attainment (master's or higher).

**Table 3. Teacher Performance Measures**

| Performance Measure | All Teachers | Math Teachers | ELA Teachers | Generalist Teachers |
|---|---|---|---|---|
| Panel A: Classroom Observations | | | | |
| FFT (Aggregate) | 2.52<br>(.316) | 2.46<br>(.329) | 2.49<br>(.364) | 2.61<br>(.217) |
| FFT (Math) | 2.52<br>(.304) | 2.46<br>(.329) | - | 2.59<br>(.261) |
| FFT (ELA) | 2.56<br>(.317) | - | 2.49<br>(.364) | 2.63<br>(.239) |
| $FFT^A$ (Aggregate) | 3.02<br>(.316) | 2.96<br>(.329) | 2.99<br>(.364) | 3.11<br>(.217) |
| $FFT^A$ (Math) | 3.02<br>(.034) | 2.96<br>(.329) | - | 3.09<br>(.261) |
| $FFT^A$ (ELA) | 3.06<br>(.317) | - | 2.99<br>(.364) | 3.13<br>(.239) |
| Panel B: Student Achievement | | | | |
| VAM (Aggregate) | 2.41<br>(.354) | 2.51<br>(.341) | 2.31<br>(.293) | 2.42<br>(.392) |
| VAM (Math) | 2.52<br>(.416) | 2.51<br>(.341) | - | 2.52<br>(.475) |
| VAM (ELA) | 2.31<br>(.363) | - | 2.31<br>(.293) | 2.32<br>(.422) |
| Panel C: Student Perceptions | | | | |
| Survey | 3.12<br>(.287) | 3.01<br>(.294) | 3.06<br>(.282) | 3.29<br>(.202) |
| Teachers | 1275 | 401 | 442 | 432 |

Notes. Mean (standard deviation) of teacher performance measures reported. Data are from the 2009–2010 school year. VAM and Survey measures have been transformed so that they are on a 1–4 continuous scale. $FFT^A$ is the teacher's adjusted FFT score, adjusted by adding 0.5 points to *FFT* score. For subject-specialists teaching multiple sections of the same subject (i.e., math or ELA teachers), aggregate FFT and VAM scores are a weighted average (weighted by section enrollment) for a single subject. For subject-matter generalists, aggregate FFT and VAM scores represent the average of math and ELA scores for the same section (class) of students. Among the full sample of 1,275 teachers, 833 teachers had VAM (math), 874 teachers had VAM (ELA), 817 had FFT (math), and 867 had FFT (ELA) scores.

**Table 4. Correlation Matrix of Teacher Performance Measures**

| | FFT[A] (Aggregate) | VAM (Aggregate) | Survey |
|---|---|---|---|
| FFT[A] (Aggregate) | 1.00 | | |
| VAM (Aggregate) | 0.11 | 1.00 | |
| Survey | 0.41 | 0.17 | 1.00 |

Notes: All correlations are statistically significant at the 0.001 level. There are 1,275 teachers in the sample.

**Table 5. Simulated Teacher Proficiency Rates, by Performance Measure Weights and District Ratings Thresholds**

| FFT (%) | VAM (%) | Survey (%) | Chicago System | Clark County System | Denver System | Fairfax County System | Gwinnett County System | Miami-Dade System | NYC System | Philadelphia System |
|---|---|---|---|---|---|---|---|---|---|---|
| **Panel A**: $Score_1$ (Survey/VAM = $\frac{1}{10}$) | | | | | | | | | | |
| 0 | 90.9 | 9.1 | .12 | .15 | .09 | .45 | .32 | .02 | .02 | .47 |
| 10 | 81.8 | 8.2 | .14 | .17 | .09 | .53 | .38 | .02 | .02 | .54 |
| 20 | 72.7 | 7.3 | .17 | .21 | .10 | .62 | .46 | .02 | .02 | .64 |
| 30 | 63.6 | 6.4 | .20 | .26 | .13 | .72 | .57 | .02 | .02 | .73 |
| 40 | 54.5 | 5.5 | .26 | .34 | .17 | .80 | .66 | .02 | .02 | .81 |
| 50 | 45.5 | 4.5 | .37 | .45 | .24 | .85 | .75 | .03 | .03 | .86 |
| 60 | 36.4 | 3.6 | .48 | .56 | .33 | .89 | .80 | .04 | .03 | .90 |
| 70 | 27.3 | 2.7 | .58 | .64 | .44 | .91 | .85 | .07 | .05 | .91 |
| 80 | 18.2 | 1.8 | .66 | .70 | .55 | .93 | .87 | .14 | .11 | .92 |
| 90 | 9.1 | 0.9 | .70 | .76 | .62 | .94 | .88 | .22 | .19 | .93 |
| 100 | 0.0 | 0.0 | .76 | .79 | .67 | .93 | .89 | .31 | .27 | .94 |
| **Panel B**: $Score_2$ (Survey/VAM = $\frac{1}{5}$) | | | | | | | | | | |
| 0 | 83.3 | 16.7 | .14 | .17 | .09 | .52 | .37 | .03 | .02 | .54 |
| 10 | 75.0 | 15.0 | .16 | .20 | .10 | .61 | .45 | .03 | .02 | .62 |
| 20 | 66.7 | 13.3 | .19 | .24 | .13 | .69 | .54 | .02 | .02 | .71 |
| 30 | 58.3 | 11.7 | .24 | .31 | .15 | .77 | .63 | .03 | .02 | .79 |
| 40 | 50.0 | 10.0 | .32 | .40 | .20 | .83 | .71 | .03 | .02 | .84 |
| 50 | 41.7 | 8.3 | .41 | .50 | .28 | .88 | .78 | .04 | .03 | .89 |
| 60 | 33.3 | 6.7 | .52 | .59 | .36 | .90 | .83 | .05 | .04 | .91 |
| 70 | 25.0 | 5.0 | .61 | .67 | .47 | .91 | .86 | .08 | .06 | .92 |
| 80 | 16.7 | 3.3 | .67 | .72 | .56 | .93 | .88 | .15 | .11 | .93 |
| 90 | 8.3 | 1.7 | .71 | .76 | .63 | .93 | .88 | .23 | .20 | .94 |
| 100 | 0.0 | 0.0 | .76 | .79 | .67 | .93 | .89 | .31 | .27 | .94 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Panel C**: $Score_3$ (Survey/VAM = $\frac{1}{2}$) | | | | | | | | | | |
| 0 | 66.7 | 33.3 | .21 | .26 | .14 | .71 | .56 | .03 | .03 | .73 |
| 10 | 60.0 | 30.0 | .25 | .31 | .16 | .78 | .63 | .03 | .02 | .80 |
| 20 | 53.3 | 26.7 | .29 | .38 | .19 | .84 | .69 | .03 | .02 | .85 |
| 30 | 46.7 | 23.3 | .37 | .47 | .23 | .87 | .77 | .03 | .02 | .88 |
| 40 | 40.0 | 20.0 | .45 | .53 | .30 | .90 | .82 | .04 | .03 | .91 |
| 50 | 33.3 | 16.7 | .53 | .61 | .39 | .92 | .85 | .05 | .04 | .92 |
| 60 | 26.7 | 13.3 | .60 | .67 | .47 | .93 | .87 | .07 | .05 | .93 |
| 70 | 20.0 | 10.0 | .65 | .71 | .53 | .93 | .88 | .11 | .08 | .94 |
| 80 | 13.3 | 6.7 | .69 | .74 | .59 | .94 | .88 | .19 | .14 | .94 |
| 90 | 6.7 | 3.3 | .72 | .77 | .64 | .94 | .89 | .25 | .21 | .94 |
| 100 | 0.0 | 0.0 | .76 | .79 | .67 | .93 | .89 | .31 | .27 | .94 |
| **Panel D**: $Score_4$ (Survey/VAM = 0) | | | | | | | | | | |
| 0 | 100 | 0.0 | .10 | .13 | .07 | .37 | .27 | .02 | .02 | .38 |
| 10 | 90 | 0.0 | .12 | .15 | .08 | .44 | .32 | .02 | .02 | .47 |
| 20 | 80 | 0.0 | .13 | .17 | .09 | .54 | .39 | .02 | .02 | .55 |
| 30 | 70 | 0.0 | .17 | .22 | .11 | .64 | .48 | .02 | .02 | .66 |
| 40 | 60 | 0.0 | .22 | .28 | .15 | .73 | .60 | .02 | .02 | .75 |
| 50 | 50 | 0.0 | .31 | .39 | .20 | .82 | .69 | .03 | .02 | .83 |
| 60 | 40 | 0.0 | .43 | .51 | .29 | .87 | .78 | .04 | .03 | .88 |
| 70 | 30 | 0.0 | .55 | .61 | .40 | .90 | .83 | .06 | .05 | .91 |
| 80 | 20 | 0.0 | .64 | .69 | .52 | .92 | .87 | .12 | .09 | .92 |
| 90 | 10 | 0.0 | .70 | .75 | .61 | .93 | .88 | .21 | .18 | .93 |
| 100 | 0 | 0.0 | .76 | .79 | .67 | .93 | .89 | .31 | .27 | .94 |

Notes. Each cell reports the proportion of teachers rated level 3 or level 4 for a given set of performance measure weights and based on the ratings thresholds of a given evaluation system . All scores are based on $FFT^A$ (Aggregate). Results in each column (within a panel) include teacher data pooled across the six MET study districts, and are based on simulations of teachers' performance scores and the ratings thresholds of district-specific teacher evaluation systems. Panel A presents the distribution of simulated teacher ratings based on a teacher's summative performance score where the ratio of the Survey/VAM weights is set to 1/10 (i.e., $Score_1$ ); Panel B presents the distribution of simulated teacher ratings based on a teacher's summative performance score where the ratio of the Survey/VAM weights is set to 1/5 (i.e., $Score_2$ ); Panel C presents the distribution of simulated teacher ratings based on a teacher's summative performance score where the ratio of the Survey/VAM weights is set to 1/2 (i.e., $Score_3$ ); and Panel D presents the distribution of simulated teacher ratings based on a teacher's summative

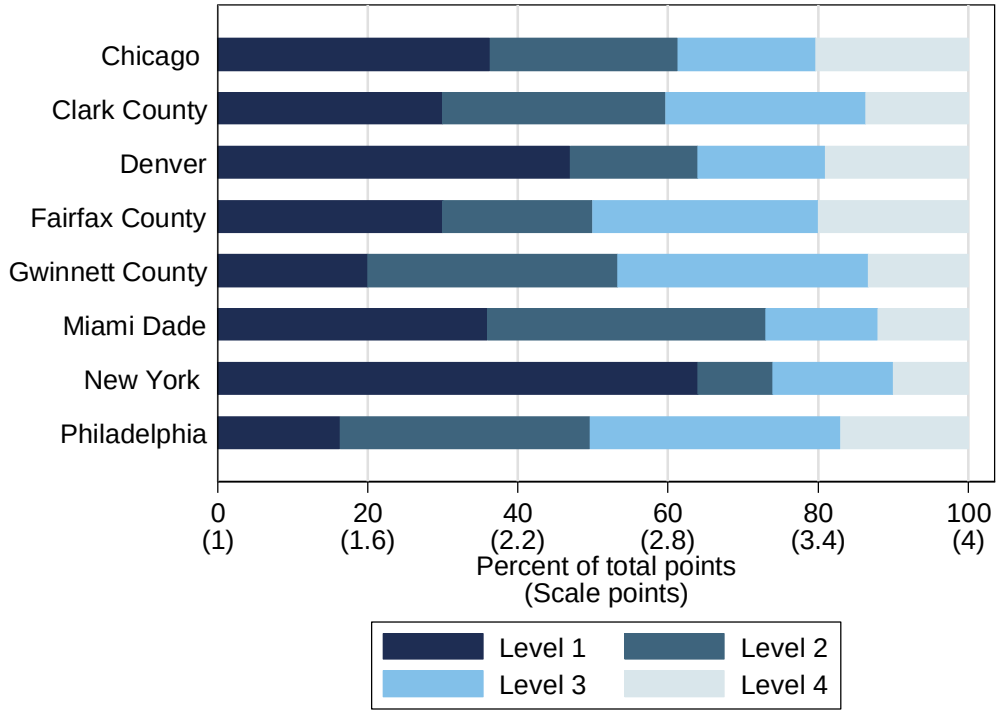
performance score where the ratio of the Survey/VAM weights is set to 0 (i.e., $Score_4$ ). See Figure 1 and Table 1 for each district's rating thresholds. There are 1,275 teachers in the sample.

**Table 6. Simulated Teacher Proficiency Rates, by Teacher Performance Score Construction (Survey/VAM Weight Ratios)**

| Teacher Performance Score (Survey/VAM Weight Ratio) | Chicago System | Clark County System | Denver System | Fairfax County System | Gwinnett County System | Miami-Dade System | NYC System | Philadelphia System |
|---|---|---|---|---|---|---|---|---|
| $Score_1$ (Survey/VAM= $\frac{1}{10}$ ) | .37 | .45 | .24 | .85 | .75 | .03 | .03 | .86 |
| $Score_2$ (Survey/VAM= $\frac{1}{5}$ ) | .41 | .50 | .28 | .88 | .78 | .04 | .03 | .89 |
| $Score_3$ (Survey/VAM= $\frac{1}{2}$ ) | .53 | .61 | .39 | .92 | .85 | .05 | .04 | .92 |
| $Score_4$ (Survey/VAM= 0) | .31 | .39 | .20 | .82 | .69 | .03 | .02 | .83 |
| Minimum | .31 | .39 | .20 | .82 | .69 | .03 | .02 | .83 |
| Maximum | .53 | .61 | .39 | .92 | .85 | .05 | .04 | .92 |
| Range | .22 | .22 | .19 | .10 | .16 | .02 | .02 | .09 |

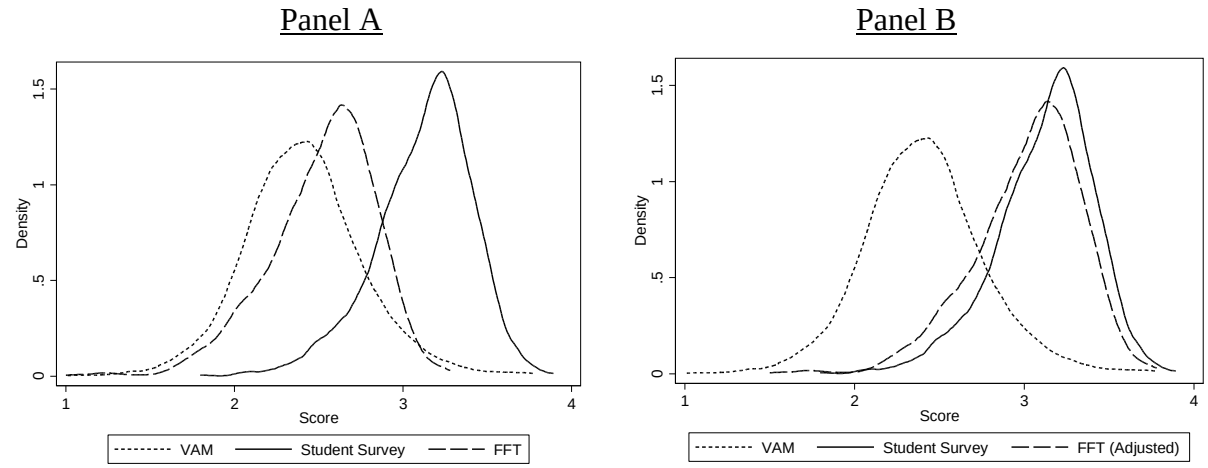
Notes. Each cell reports the proportion of teachers rated level 3 or level 4 for a given construction of teacher's summative performance scores. Results are based on weights which fix $FFT^A$ (Aggregate) at 50%, and allow Survey and VAM weights to vary. $Score_1$ is based on a Survey/VAM weights ratio of 1/10; $Score_2$ is based on a Survey/VAM weights ratio of 1/5; $Score_3$ is based on a Survey/VAM weights ratio of 1/2; and $Score_4$ is based on a Survey/VAM weights ratio of 0. There are 1,275 teachers in the sample.

**Figure 1. Evaluation System Ratings Thresholds**




Notes. Stacked bars represent the range of evaluation score points associated with each of the four different evaluation rating categories across eight districts. Level 1 corresponds with a district's lowest (of four) evaluation rating; level 3 corresponds with a "Proficient" or "Effective" rating; level 4 corresponds with a district's highest evaluation rating. Performance rating thresholds for each district's evaluation system are captured by the vertical lines separating rating category. All scoring systems have been rescaled so that they map onto a common point scale ranging from 1 to 4 total available evaluation system points. See Table 1 for more detail on each district's evaluation system.

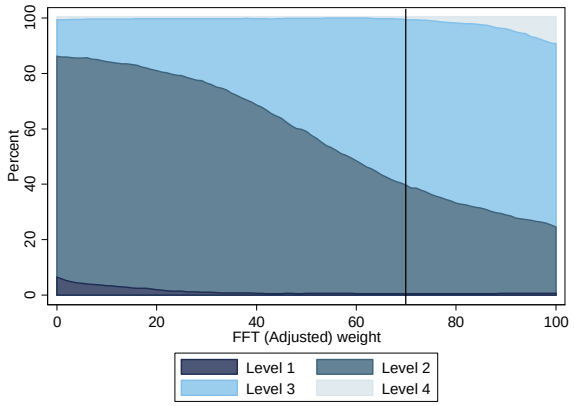
**Figure 2. Distribution of Teacher Performance Scores**



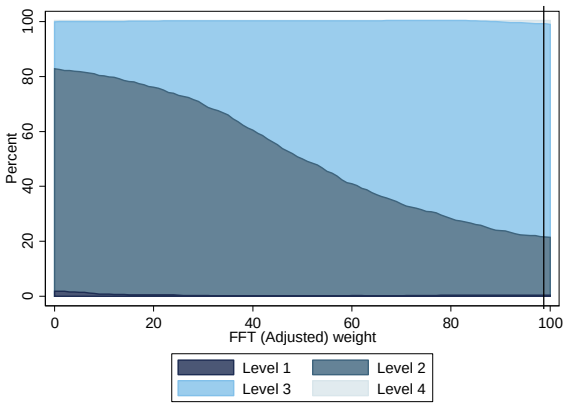

Notes. Data are from the 2009–2010 school year. Sample includes all teachers (n=1,275). In Panels A and B, VAM is the VAM (Aggregate) measure from Table 3; *Survey* is the Survey measure from Table 3. In Panel A, *FFT* is the FFT (Aggregate) measure from Table 3. In Panel B, *FFT (Adjusted)* is the $FFT^{A}$ (Aggregate) from Table 3.

**Figure 3. Distribution of Teacher Ratings, by Performance Measure Weights and System Ratings Thresholds**

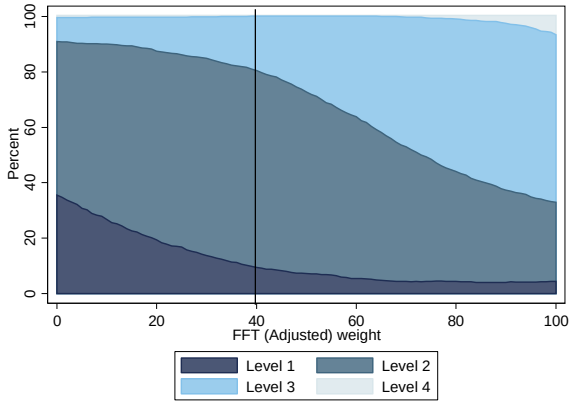
Panel A: Chicago



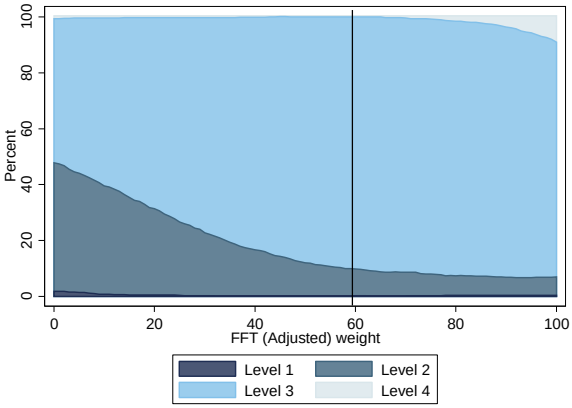

Panel B: Clark County



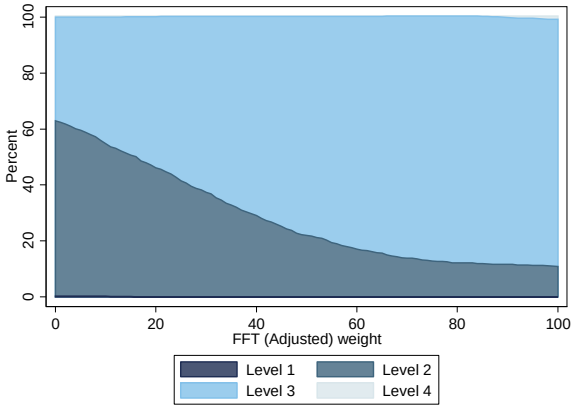

Panel C: Denver



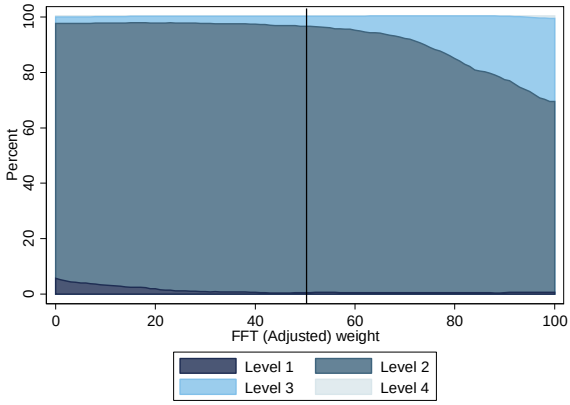

Panel D: Fairfax County



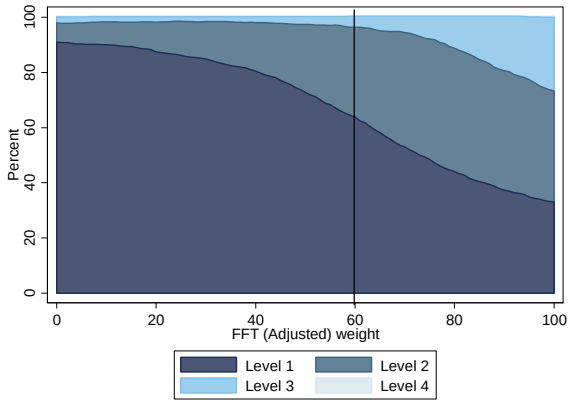

Panel E: Gwinnett County



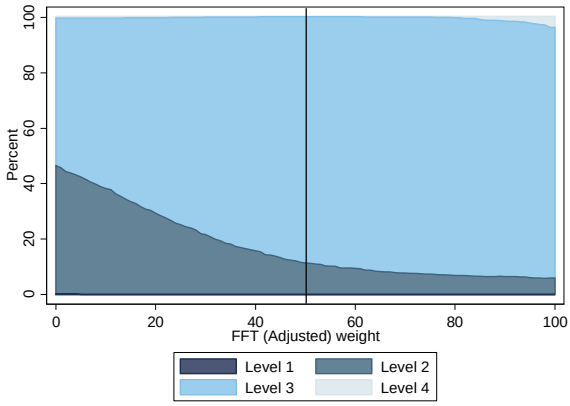

Panel F: Miami-Dade County



Panel G: New York City

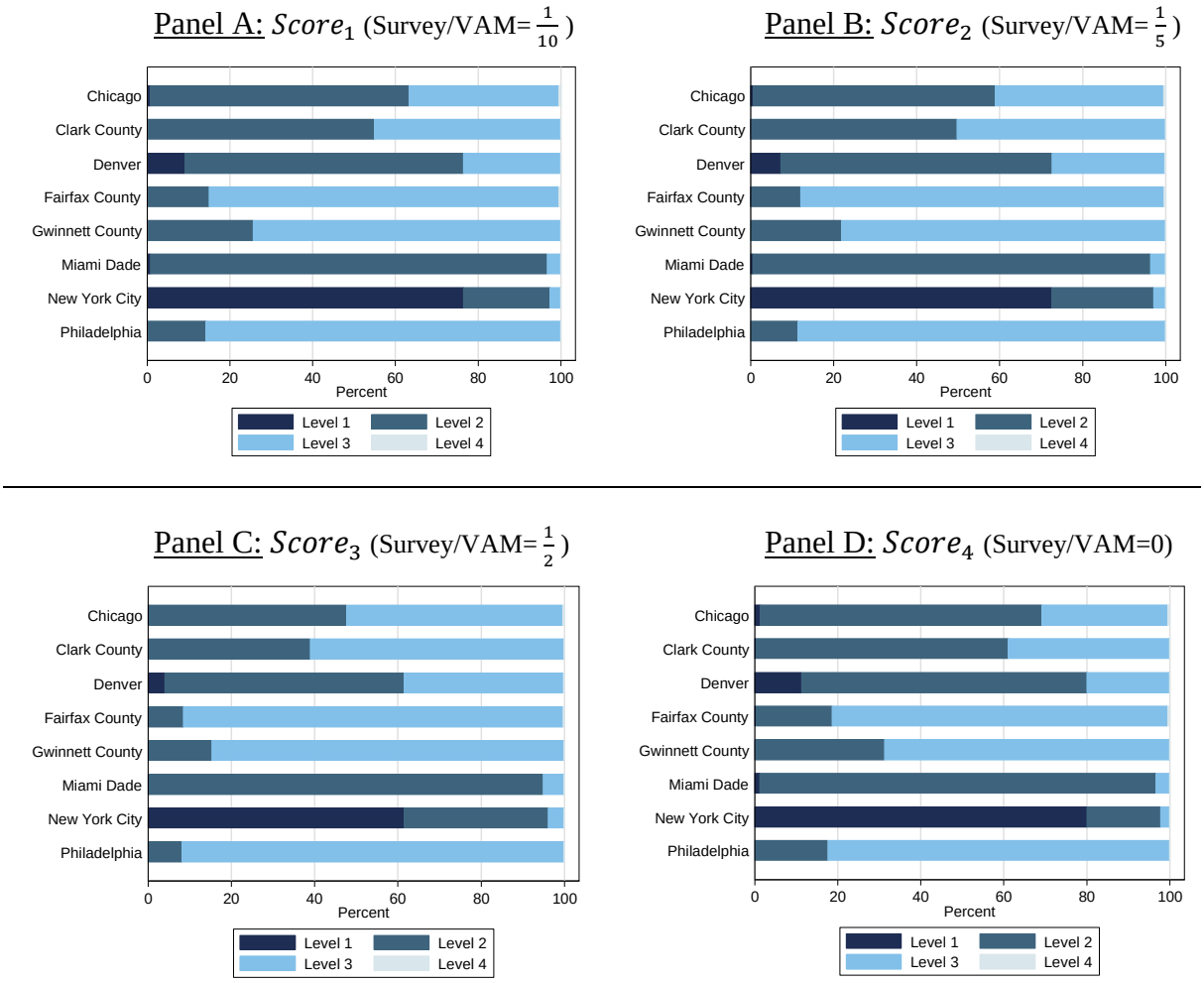



Panel H: Philadelphia




Notes. Each panel shows the distribution of teacher ratings (levels 1–4) based on a given evaluation system's performance rating thresholds and across different weighting schemes assigned to a teacher's summative evaluation score with the Survey to VAM ratio of 1/5. Teacher summative evaluation scores based on weights assigned to observation score (*FFT*$^{A}$ *(Aggregate)*), VAM score (*VAM (Aggregate)*) and survey score (*Survey*) (see Table 3). Sample includes all teachers (n=1,275). The vertical line indicates the weight assigned to observation scores in each district's evaluation system (see Table 1).

**Figure 4. Distribution of Teacher Ratings, by System Ratings Thresholds and Teacher Performance Score Construction (Survey/VAM Weight Ratios)**




Notes. Each panel shows the distribution of teacher ratings (levels 1–4) based on a given evaluation system's performance rating thresholds and one (of four) score constructions. A teacher's summative evaluation score is based on weights assigned to observation score (*FFT*$^A$ *(Aggregate)*), VAM score (*VAM (Aggregate)*) and survey score (*Survey*) (see Table 3). In Panel A, a teacher's summative evaluation score (*Score 1*) is based on $Ratio_{Survey/VAM} = \frac{1}{10}$ and the following performance measure weights: FFT=50%; VAM=45.5%; and Survey=4.5%. In Panel B, a teacher's summative evaluation score (*Score 2*) is based on $Ratio_{Survey/VAM} = \frac{1}{5}$ and the following performance measure weights: FFT=50%; VAM=41.7%; and Survey=8.3%. In Panel C, a teacher's summative evaluation score (*Score 3*) is based on $Ratio_{Survey/VAM} = \frac{1}{2}$ and the following performance measure weights: FFT=50%; VAM=33.3%; and Survey=16.7%. In Panel D, a teacher's summative evaluation score (*Score 4*) is based on $Ratio_{Survey/VAM} = 0$ and the following performance measure weights: FFT=50% and VAM=50%. Sample includes all teachers (n=1,275).

## Data Appendix

### Chicago, IL

Chicago Public Schools. (2014). REACH Students: Educator Evaluation Handbook 2014-2015. p.61. Retrieved from: http://www.ctunet.com/rights-at-work/teacher-evaluation/text/CPS-REACH-Educator-Evaluation-Handbook-FINAL.pdf.

### Clark County, NV

Nevada State Board of Education. (2015, September). NEPF Educator Performance Framework (NEPF): Statewide Evaluation System. p.17. Retrieved from[1]: http://www.doe.nv.gov/Educator_Effectiveness/Educator_Develop_Support/NEPF/Tools_and_Protocols/

### Denver, CO

Denver Public Schools. (n.d.). LEAP Handbook 2014-2015. p.5. Retrieved from: http://www.nctq.org/docs/denver_2014-15-LEAP-handbook-master_1.pdf.

### Fairfax County, VA

Fairfax County Public Schools. (2015, August). Teacher Performance Evaluation Program Handbook. p.16. Retrieved from: http://www.nctq.org/docs/TEHandbook.pdf

### Gwinnett County, GA

Gwinnett County Public Schools. (n.d.). A Primer For Teachers 2015-2016. p.6. Retrieved from: http://www.nctq.org/docs/2015-16-GTES-Primer_FINAL_June25.pdf

### Miami-Dade, FL

Miami-Dade County Public Schools. (2015). IPEGS Procedural Handbook 2015 Edition. p.92. Retrieved from http://ipegs.dadeschools.net/pdfs/2015_IPEGS_Procedural_Handbook.pdf.

### New York City, NY

New York City Department of Education. (2014, September 17). *Advance* Overall Ratings Guide. p.12. Retrieved from http://www.uft.org/files/attachments/advance-ratings-guide-2013-14.pdf.

### Philadelphia, PA

Pennsylvania Department of Education. (2014, July). Educator Effectiveness Administrative Manual. p.19. Retrieved from http://www.nctq.org/docs/Educator_Effectiveness_Administrative_Manual.pdf

[1] After following the link, you will be directed to a page on the Nevada DOE website that gives an overview of the NEPF. To access the Nevada Educator Performance Framework document, click on the hyperlink that says "Protocols" under the headline "NEPF Protocols" to download the document.