NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

Combining Multiple Performance Measures: Do Common Approaches Undermine Districts' Personnel Evaluation Systems?

MICHAEL HANSEN
MARIANN LEMKE
NICHOLAS SORENSEN

WORKING PAPER 118 • OCTOBER 2014

Combining Multiple Performance Measures: Do Common Approaches Undermine Districts' Personnel Evaluation Systems?

Michael Hansen
American Institutes for Research

Mariann Lemke
American Institutes for Research

Nicholas Sorensen
American Institutes for Research

Contents

Acknowledgements

This research was supported by the National Center for Analysis of Longitudinal Data in Education Research (CALDER) funded through Grant R305C120008 to the American Institutes for Research from the Institute of Education Sciences, U.S. Department of Education.

CALDER working papers have not undergone final formal review and should not be cited or distributed without permission from the authors. They are intended to encourage discussion and suggestions for revision before final publication.

The views expressed are those of the authors and should not be attributed to the American Institutes for Research, its trustees, or any of the funders or supporting organizations mentioned herein. Any errors are attributable to the authors.

The authors gratefully acknowledge Tiffany Chu of AIR for providing excellent research assistance and Cory Koedel (University of Missouri), Dan Goldhaber (University of Washington–Bothell), and Mike Garet and Dan Sherman of AIR for helpful comments.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

Combining Multiple Performance Measures: Do Common Approaches Undermine Districts' Personnel Evaluation Systems?

Michael Hansen, Mariann Lemke, and Nicholas Sorenson

CALDER Working Paper No. 118
October 2014

Abstract

Teacher and principal evaluation systems now emerging in response to federal, state and/or local policy initiatives typically require that a component of teacher evaluation be based on multiple performance metrics, which must be combined to produce summative ratings of teacher effectiveness. Districts have utilized three common approaches to combine these multiple performance measures, all of which introduce bias and/or additional prediction error that was not present in the performance measures originally. This paper investigates whether the bias and error introduced by these approaches erodes the ability of evaluation systems to reliably identify high- and low-performing teachers. The analysis compares the expected differences in long-term teacher value-added among teachers identified as high- or low-performing under these three approaches, using simulated data based on estimated inter-correlations and reliability of measures in the Gates Foundation's Measures of Effective Teaching project. Based on the results of our simulation exercise presented here, we conclude these approaches can undermine the evaluation system's objectives in some contexts. Depending on the way these performance measures are actually combined to categorize teacher performance, the additional error and bias can be large enough to undermine the district's objectives.

Introduction

States and districts across the country currently are overhauling their teacher and principal evaluations systems to promote effective teaching and leadership in schools. Previously, teacher evaluation systems made few qualitative distinctions among teachers, with 94 percent of teachers considered effective or better (Weisberg, Sexton, Mulhern, & Keeling, 2009). Recent national policy initiatives, including Race to the Top and the Elementary and Secondary Education Act (ESEA) flexibility waivers for states, have provided a major impetus for reform, requiring states and districts to make significant changes to their teacher and principal evaluation systems on rapid timelines in exchange for funding or flexibility regarding specific requirements of the 2002 reauthorization of ESEA. The primary goal of these reforms is to create evaluation systems that better differentiate effective and ineffective teaching and leadership so that districts and schools can make more informed personnel decisions.

The teacher and principal evaluation systems emerging in response to federal, state, or local policy initiatives typically require that a significant component of the evaluation be based on student performance in conjunction with other performance measures. To measure student performance, many states and districts are adopting value-added models, which provide a statistical estimate of a teacher's or principal's contribution to student learning. Other performance measures may include observation-based measures of practice, parent or student survey results, and measures of professionalism, among others. Districts are then tasked with combining value-added estimates with these other disparate measures to produce a single summative rating of effectiveness for each teacher or principal. In doing so, the designers of these evaluation systems must make careful decisions not only about selecting valid and reliable measures but also about how to combine performance ratings to minimize error and misclassification.

To date, the burden of how to integrate these multiple measures into a single performance rating has largely fallen on states and districts. Three commonly used models have emerged in practice, all of

which (unintentionally) introduce error or bias that did not exist in the performance measures originally. This paper, written by researchers at American institutes for Research (AIR), investigates whether the bias and error introduced by these approaches erode the ability of evaluation systems to reliably identify high- and low-performing teachers, based on their long-term value-added productivity. In summary of our findings, we conclude one of the common approaches—the numeric model—is the preferred approach across a variety of contexts and in many cases is not statistically different from an optimal approach that cannot be implemented in practice. Also, we find that the other two common approaches can undermine the evaluation system's objectives in some contexts.

In the section titled The District's Evaluation Problem, we describe the district's problem in attempting to evaluate a teacher's performance using measures that are fraught with error. The Simulation section describes the simulation that we conduct, based on estimated intercorrelations and reliability of measures in the Bill & Melinda Gates Foundation's Measures of Effective Teaching (MET) project. The Results section presents the results of these simulations, and the Conclusion and Discussion section provides recommendations.

The District's Evaluation Problem

A growing body of evidence shows teachers are the most consequential component of schooling, contributing to gains in student learning (as measured by standardized tests), and principals are the second most important input (e.g., Branch, Hanushek, & Rivkin, 2009; Hanushek & Rivkin, 2010; Staiger & Rockoff, 2010). Districts that reliably identify high- and low-performing teachers and principals, as measured through long-term value-added performance, and use this information in personnel management (for example, through compensation, selective retention) can thereby improve student learning by having the most productive teachers and principals staff their schools.

Unfortunately, identifying high and low performers is easier said than done. Easily verifiable and observable characteristics are not well correlated with performance, and a teacher's or principal's performance may fluctuate over time, making long-term value-added difficult to reliably infer (Aaronson, Barrow, & Sander, 2007; Goldhaber & Hansen, 2013). The best predictors of current value-added effectiveness are past value-added estimates, but these estimates are prone to considerable measurement error (Goldhaber & Hansen, 2010; McCaffrey, Sass, Lockwood, & Mihaly, 2009). In addition, some researchers have criticized an overreliance on value-added estimates for a variety of reasons, including the lack of teacher support and their inability to be calculated for teachers outside of tested grades and subjects (Baker et al., 2010).

Given the practical and political difficulties of relying on value-added measures exclusively to evaluate teachers' performance, districts and states have instead shifted toward using multiple performance measures to build these emerging evaluation systems. Three reports from the Gates Foundation's MET project (hereafter referred to collectively as "MET studies") have been the most authoritative resources on the interrelationship between these various performance measures (Bill & Melinda Gates Foundation, 2010, 2012, 2013).[1] These studies produce several performance measures for participating teachers, including value-added estimates, student survey scores, and several different observation-based ratings. In general, they find low, positive correlations across these different performance metrics overall. The 2012 MET study interprets this evidence to suggest that the disparate performance measures may be jointly used to better estimate long-term teacher value-added performance, which it refers to as "underlying value added." [2]

[1] Other studies also have investigated the correlation between various types of teacher performance measures, including Grossman et al. (2010), Jacob and Lefgren (2008), and Rockoff and Speroni (2010).

[2] Long-term value-added is the theoretical averaged value-added productivity of teachers if one could observe value-added performance over a teacher's entire career. Note that this long-term value-added is a theoretical concept and is not formally measured for any individual teacher in the 2012 MET study; rather, the correlations of underlying teacher quality with the three performance measures are derived using true-score theory.

The district's evaluation problem,[3] therefore, is to jointly use these various performance measures on its teachers and principals to identify the highest and lowest performers based on long-term value-added productivity, which cannot be observed directly.[4] In other words, the district must use these imperfect measures on multiple dimensions to minimize misclassification error across different effectiveness categories. The process of combining these measures may take one of several forms, though this choice may influence how successful the evaluation system is in identifying high and low performers. Our primary research question is how this choice of approach in combining multiple performance measures will influence an evaluation system's ability to reliably identify high- or low-performing staff members.

Consider a district's evaluation system that collects performance measures for teachers from three different sources: value-added estimates, external observations, and a student survey.[5]
The district then must use these three measures to classify its teachers into one of four distinct effectiveness categories (in ascending order): ineffective (IE), marginally effective (ME), effective (E), and highly effective (HE). Assume that the district intends to use these four categories to identify the teachers that it infers to come from various segments of the distribution of the target criterion, long-term value-added performance: IE captures the bottom 10 percent of the distribution, ME captures the 11th to 20th percentile, E captures the 21st to 80th percentile, and HE captures the top 20 percent of the distribution. Though the particulars vary, this basic framework is representative of the evaluation

[3] In framing this discussion and in the simulation results that follow, we refer to this evaluation problem as the district's responsibility; but in practice, states also may make policy that shapes personnel evaluation systems.
[4] Note that the target criterion used here is long-term value-added, which is the same target criterion used in most of the MET studies and simplifies the analysis, although other target criteria could be substituted. A companion research paper (Mihaly, McCaffrey, Staiger, & Lockwood, 2013) to the 2013 MET study uses the project data to investigate how the multiple measures may be combined using various other target criteria, including maximizing all dimensions of measured teacher performance jointly. Because the main research question in our study is how the method of combining these disparate measures influences who is identified as high or low performing, investigating other target criteria is beyond the scope of this paper.
[5] The remainder of the paper analyzes the problem in the context of a district evaluating the teacher workforce; applying these findings to the case of evaluating principals is directly analogous.

systems that have emerged in front-runner districts on teacher evaluation reform and is adopted in the simulation presented later.[6]

The district's most efficient option (i.e., that which minimizes the mean squared error on the district's prediction on the target criterion) is to use the correlations of the performance metrics—both across measures and within measures over time—to infer the optimal weighting across measures in predicting long-term value-added performance. Implicitly, this approach creates a single combined measure that is a weighted average of the component measures, where the weights are determined based on the empirical properties of the component measures. The correlations of the various metrics with long-term value added derived in the 2012 MET study using true-score theory provides one method to uncover the weights in this relationship.[7] Alternatively, one may consider the optimal weights as the estimated coefficients in a theoretical regression (if one could measure long-term value-added performance) using the three component measures to predict long-term value-added performance.[8] This regression has its analogues in other research papers that use regression-based methods to predict teachers' future performance, on the premise that future performance approximates long-term performance (e.g., Goldhaber & Hansen, 2010; Lefgren & Sims, 2012; Rockoff & Speroni, 2010).

Based on teachers' predicted values of long-term value-added performance, districts can then assign summative effectiveness categories to those teachers based on where they fall in the predicted

[6] For example, the teacher evaluation systems implemented in Denver, Hillsborough County (Florida), Pittsburgh, and Washington, D.C., all have evaluation systems similar to that used here (where there are two levels of low performance intended to capture the lowest part of the teacher distribution, one level of high performance capturing the top of the teacher distribution, and a large average-performance category that captures most teachers in the district). However, these districts vary on the criteria for earning each category designation and the proportion of teachers that fall into each group. Memphis (Tennessee) and New Haven (Connecticut) both have five-category evaluation systems that have two small high-performing categories, which mirror the two small low-performing categories, and still include one large average category intended to capture most teachers, similar to that seen in the four-category districts listed earlier.

[7] True-score theory is a psychometric method to uncover latent traits using multiple measures. The 2012 MET study assumes the errors across the three types of measures included in the study are independent, and the long-term value-added performance of teachers is constant over time. See the 2012 MET study for further details.

[8] This regression approach is the method we use to implement the error-minimizing approach presented in the results tables. The estimated beta coefficients for each of the three components are presented in the Appendix.

distribution. Although this method minimizes the error in predicting long-term performance (and is hereafter referred to as the "error-minimizing approach"), it is not used in formal evaluation systems because it removes the decision of weighting the various performance measures from the state or district.

Three Methods to Combine Multiple Performance Measures

In practice, three commonly used methods have emerged as districts have formally adopted multiple performance measures into their evaluation systems. These approaches are referred to in Leo and Lachlan-Haché (2012) as the *numeric* approach, the *hybrid* approach, and the *profile* approach.[9] As will be described, all three of these methods introduce error (decreasing the ability to accurately infer long-term value-added performance), and two of them also introduce bias (systematically rating teachers higher or lower than they really are) even if the source measures themselves contain no bias. Consequently, districts' use of these approaches to combine multiple performance measures may potentially undermine the intended objective of reliably identifying high and low performers in the district. Whether the shortcomings of these approaches are severe enough to render the evaluation systems ineffectual is an empirical question that we address here through a simulation.

All three of the commonly used approaches share a similar departure from the error-minimizing approach described earlier, which is the determination of weights for each of the component performance metrics through some external means.[10] Whether set by negotiations with the teachers union, district policy, or state legislation, any external process to determine the weights will inflate the importance of some performance components at the expense of others in predicting the target

[9] These three approaches are described in detail in Leo and Lachlan-Haché (2012), along with illustrative examples of their implementation and documentation of several districts that have adopted these various approaches.

[10] Any discussion of weighting the different component measures requires that the raw scores of any measurement type be standardized first before combining with other metrics. If the variances were not equalized, the measure with the largest spread would have the largest implied weight in the final score, regardless of whether that was intended. The MET studies similarly standardize the performance measures to investigate weighting issues.

criterion, thereby increasing error in the district's prediction of long-term value-added performance. Yet, as long as none of the performance metrics themselves contains bias, the introduction of these external weights will increase error only and will not introduce bias. In our simulation, we use weights of 50 percent on the value-added estimates, 35 percent on the observation scores, and 15 percent on the student survey scores as the starting point, although we also will investigate how varying these weights influence the identification of high and low performers.[11] Given these weights, the district then chooses among the three commonly used approaches to combine these performance metrics, each of which is described in more detail below (see Leo & Lachlan-Haché, 2012, for a lengthier discussion of each approach).

Numeric Approach

First, a *numeric approach* is most similar to the error-minimizing approach, with the only difference being the use of the external weights rather than empirically determined weights for the performance metrics. These weights are applied directly to the standardized performance measures, forming a single weighted average across all types of measures. A teacher's summative effectiveness rating is then a function of where the teacher falls in the distribution of this combined measure. Because this approach creates a weighted average of continuous measures, the individual measures are compensatory, where high performance in value-added, for example, can directly substitute for low performance on student surveys. Aside from the use of external weights, which introduce error in predicting the target criterion, this method is the nearest match to the error-minimizing approach and does not bias the inference of long-term value-added performance.

Hybrid Approach

[11] This weighting is on the high end of how states and districts weight value-added estimates in practice, although we choose this weighting to start because it is both representative of some districts' weighting schemes and is relatively well aligned with the underlying reliabilities of the performance measures on long-term value-added performance. See the estimated coefficients of the error-minimizing approach in the Appendix for comparison.

Next, the *hybrid approach* is similar to the numeric approach described earlier but reorders the steps in the process. In the hybrid approach, teachers' performance in each dimension is first categorized based on where they fall in the distribution of that particular metric, resulting in three separate ratings for each teacher corresponding to each metric that takes on categorical values. These categorical values are mapped onto integer values (e.g., an IE rating = 1, ME = 2), and then weights are applied to these integer values to create an overall score that will be used to infer the target criterion.[12] Summative ratings are determined by rounding the overall score and mapping the integer values back to effectiveness ratings to which the integers are equivalent.

Though the reordering sounds innocuous, it adversely affects the prediction on the target criterion in two ways. First, the hybrid approach increases the error on the overall combined score because the categorizations on each measure ignore variations within each effectiveness category before combining the measures (e.g., a teacher in the 25th percentile would be considered equivalent to a teacher in the 75th percentile) when it could have been used to improve the predictive power on the combined measure. Second, assigning integer values to the category ratings and then rounding implicitly introduces a small bias that favors teachers. This situation occurs whenever the categorical values of the component metrics disagree and result in a combined weighted score in the middle between measures and are subsequently rounded to create the final summative rating (e.g., 3.5, which is rounded to 4 = HE). Rounding to the higher number systematically favors teachers in the event of such ties, which are expected to occur with some frequency given that all component measures are reduced to integer values (exact ties rarely occur with continuous measures).[13]

---

[12] This categorization of performance by type before combining implies that the performance measures are no longer directly compensatory, capturing only the differences resulting from moving across the discrete thresholds for performance.

[13] Alternatively, some systems may round downward or to the average category when different metrics disagree about a teacher's performance. This decision rule will similarly introduce a small bias, although the direction of the bias will change from what we estimate here.

Profile Approach

Finally, the *profile approach,* like the hybrid approach, categorizes teacher performance along each measure first before combining, but it differs in that it combines the component measures in multiple steps rather than in a single weighted average calculation. Commonly, evaluation systems that use the profile approach will present a series of decision matrices that determine a teacher's overall rating, given his or her respective effectiveness ratings on each component measure.[14]

Figure 1 presents a simple example of how this approach might be implemented with three component measures. The student survey and observation measures are combined in Step 1 to result in a single combined rating for both. The combined rating from Step 1 is then combined with the value-added rating to arrive at a final overall summative effectiveness rating in Step 2. By combining integer values and rounding in multiple steps, the profile approach increases the prediction error on the resulting combined score and further increases the potential for bias that systematically favors teachers.

[14] When using the decision matrices common to this approach, users do not explicitly observe the weights that determine the resulting classifications; however, the decision profiles are designed in such a way as to reflect an underlying weighting scheme across different types of performance measures.

Figure 1. Typical Decision Tables Under the Profile Approach

Step 1. Finding the Survey-Observation Combined Rating

		Student Survey Rating			
		IE	ME	E	HE
Observation Rating	HE	E	E	HE	HE
	E	ME	E	E	E
	ME	ME	ME	ME	E
	IE	IE	IE	ME	ME

Key

IE Ineffective

ME Marginally effective

E Effective

HE Highly effective

Step 2. Finding the Overall Summative Rating

		Value-Added Rating			
		IE	ME	E	HE
Survey-Observation Rating	HE	E	E	HE	HE
	E	ME	E	E	HE
	ME	ME	ME	E	E
	IE	IE	ME	ME	E

Figure 2 summarizes the main steps of the three competing approaches described here and provides an example of how the summative effectiveness rating under each is determined for a hypothetical teacher. This hypothetical teacher has value-added, observation, and student survey scores that fall in the 12th, 23rd, and 16th percentiles, respectively. As illustrated, starting with the same three performance measures, the three approaches all contain the same two processes of creating a combined score and categorizing performance but vary slightly in implementation.

As a result of these slight differences in their approaches, these three methods result in three different effectiveness ratings for the teacher. This particular example teacher is atypical because most teachers' classifications are not very sensitive to the approach used; and of those identified with differing classifications, the very large majority will be categorized into only two different categories—not three as shown here. Yet, this example illustrates the potential errors that are dependent on the choice of aggregation approach.

Figure 2. Illustrative Example of Computing Effectiveness Ratings Under Competing Approaches

Teacher judged to have following performance measures:

- Value-added estimate at the 12th percentile (z = −1.20), given 50% weight
- Observation ratings at the 23rd percentile (z = −0.75), given 35% weight
- Student survey results at the 16th percentile (z = −1.01), given 15% weight

Numeric Approach	**Hybrid Approach**	**Profile Approach**
1. Calculate Weighted Average on standardized performance metrics $(0.5) * (-1.20)$ $+(0.35) * (-0.75)$ $+(0.15) * (-1.01)$ $= -1.01$	1. Categorize Performance on each separate metric, map to integers VA: 12^{th} pctle $=$ ME $= 2$ OBS: 23^{rd} pctle $=$ E $= 3$ SS: 16^{th} pctle $=$ ME $= 2$	1. Categorize Performance on each separate metric, map to integers VA: 12^{th} pctle $=$ ME $= 2$ OBS: 23^{rd} pctle $=$ E $= 3$ SS: 16^{th} pctle $=$ ME $= 2$
2. Categorize Performance based on combined score distribution $-1.01 = 8^{th}$ pctle = IE	2. Calculate Weighted Average of integer values and round $(0.5) * 2$ $+(0.35) * 3$ $+(0.15) * 2$ $= 2.4$ rounds to $2 =$ ME	2. Calculate Weighted Average of integer values and round in multiple steps Step 1: $(0.70) * 3 + (0.30) * 2$ $= 2.7$ rounds to $3 =$ E Step 2: $(0.5) * 3 + (0.5) * 2$ $= 2.5$ rounds to $3 =$ E

Our investigation here examines the extent to which the introduction of the error and bias in these various approaches to combining multiple measures of teacher performance undermines the overall evaluation system. We know that at minimum, all evaluation systems that are implemented in practice are not statistically efficient (i.e., do not minimize the mean squared error in the prediction) by virtue of using externally determined weights rather than weights that are empirically determined relative to the system's target criterion. But of the three commonly used approaches, the numeric approach should be the preferred option from a measurement perspective. Yet, districts may choose one of the other two approaches for other considerations (e.g., ease of interpretation for teachers, interest in setting minimum performance criteria on particular dimensions). However, how this choice affects the district's ability to reliably identify its high- and low-performing teachers in practice is an empirical question and is unclear in the absence of a simulation. The simulation analysis addresses these issues.

We can predict how the error and bias introduced in these approaches will affect our resulting categorization of teachers. Increasing the prediction error, as these three commonly used approaches do, we expect the evaluation system overall will be relatively less reliable in identifying high- or low-performing teachers. This lower reliability means that more teachers will be misclassified overall with each incremental addition of prediction error, and the differences in true long-term value-added performance between the groups identified as high and low performers will decrease. In addition, the bias favoring teachers (introduced in the hybrid and profile approaches) will systematically increase the likelihood that a teacher is identified as more effective than he or she truly is. Although we can predict how these various approaches will differ, we cannot say whether the magnitude of these differences will be consequential or statistically significant in practice without a simulation.

Simulation

The relative performance of these competing approaches in identifying teachers as high or low performers can be readily analyzed with a simulation. The key advantage of simulated data is that we can observe the target criterion, long-term value-added (hereafter LTVA) performance for teachers, where in practice this measure is unobservable. Using the MET studies' estimated intercorrelations of the various performance measures, we randomly simulate a sample of 500 teachers' joint performance scores along with their LTVA. We then inspect how well each of the three approaches used in practice, along with the error-minimizing approach, compare in identifying teachers' LTVA over 1,000 iterations of the simulated data. The particulars of the simulation are described here.

To begin, we use the correlation estimates of teacher performance measures derived from the 2010 and 2012 MET studies as our parameter values. The MET studies investigate three yearly teacher performance measures: value-added estimates in multiple subjects, external observation ratings using multiple rubrics, and Tripod Project student surveys. These measures are all correlated in various ways, both within measures (e.g., the correlation of student survey responses across classes for a given teacher within a year) and across measures (e.g., correlating teachers' value-added estimates with external observation ratings). We use the studies' highest reported correlation estimates across each type of measure, which are most commonly observed in mathematics classrooms, to represent a best-case scenario of correlation between measurement types. The 2012 MET study also derives the correlation of all three types of performance measures against a theoretical construct of LTVA. These correlation parameters are used to jointly generate four variables (with mean 0 and standard deviation of 1) representing these three yearly performance measures and the target criterion for 500 teachers in a single iteration sample.

The target criterion, LTVA, is then used to segment the distribution of teachers into their "true" effectiveness categories, mirroring those described earlier. Ineffective teachers comprise the bottom 10

percent of teachers based on LTVA, minimally effective teachers are those in the 11th to 20th percentile, effective teachers are those in the 21st through 80th percentile, and highly effective teachers are those in the top 20 percent of the distribution.

The generated teacher performance measures are then combined using the four aggregation approaches presented earlier (the error-minimizing approach in addition to the numeric, hybrid, and profile approaches) in the district's attempt to infer teacher performance on the same target criterion. Teachers are sorted into the same four effectiveness categories (IE, ME, E, and HE), intended to represent the same part of the distribution of LTVA. The purpose of the simulation is to investigate the agreement between the true categorization (based directly on the simulated values of LTVA) and those predicted from the different aggregation approaches (based on inferences using observed teacher performance metrics).

At the conclusion of this aggregation and classification process for each iteration, we record several measures of the commonality between each approach's predicted effectiveness rating and the true categorizations for all teachers in the simulated workforce. These measures will help us in evaluating how well each of these models discriminate on LTVA, as intended. Note that several of the metrics presented here focus on the tails of the distribution (i.e., those not classified as the middle "effective" category). The objective of teacher evaluation systems like those presented here is to combine multiple measures of teacher performance to discriminate teachers among various levels of quality, particularly separating out high and low performers from the rest of the workforce. The metrics that focus on the tails, therefore, are relevant in comparing the accuracy of these competing models in categorizing teacher performance. By summarizing these measures over the 1,000 iterations in the simulation, we can understand how these different measures are expected to vary in practice. We report several measures:

- **Percentage correct by rating**—This metric reports the percentage of simulated observations each model identifies as IE, ME, E, or HE, in which the underlying true categorization agrees with this classification. Higher values indicate a better fit.
- **Percentage misclassified by rating**—This metric reports the percentage of simulated observations that each model identifies as IE, ME, E, or HE, in which the underlying true categorization disagrees with this classification. Lower values indicate a better fit.
- **Overall percentage correctly identified**—This metric reports the total percentage of teacher classifications in the sample (combining across effectiveness ratings) where the predicted rating agrees with the true rating. Higher values indicate a better fit.
- **Percentage correct in the tails**—This metric reports the percentage of observations where the predicted rating agrees with the true rating when the model assigned the teacher to the IE, ME or HE rating (i.e., removes all teachers that are identified as E, and then calculates the percentage correct among the remainder). Higher values indicate a better fit.
- **Extreme error rate**—This metric reports the percentage of observations where the predicted rating was in the lower tail (IE or ME) and the true rating was in the upper tail (HE) or vice versa. Lower values indicate a better fit.
- **Average long-term value-added of effective or better teachers**—This metric reports the average LTVA (in standard deviation units of LTVA) of teachers identified as either E or HE (i.e., removes those identified in the lower tail). Higher values indicate a better fit.
- **Difference in long-term value-added between the tails**—This metric reports the difference in group means (in standard deviation units of LTVA) between teachers identified in the lower tail (IE or ME) against those identified in the upper tail (HE). Higher values indicate a better fit.
- **Ratio of overstating to understating effectiveness**—This metric reports the ratio of two probabilities: The numerator is the probability that teachers are identified in a higher category than they actually are, and the denominator is the probability that teachers are identified in a lower category than they actually are. We take this ratio as a measure of bias, where a value of 1 indicates no bias, and values greater than 1 (or less than 1) indicate a bias favoring teachers (or penalizing teachers).

Finally, we conduct a series of variations in the simulation to investigate how these changes may affect our results. In the Results section, we report on two particular variations. The first investigates how using a different set of weights influences the differences between models. The second investigates how using reliability-adjusted thresholds to classify teachers into performance categories affects the resulting teacher categorizations. These investigations will be described in further detail along with their accompanying results in the Results section.

Discussion of Limitations and Generalizability

We need to note several important caveats about this simulation framework before proceeding to the results because they pertain to the generalizability of our findings. First, our use of these three particular performance measures (value-added estimates, observation ratings, and student surveys) is

not intended to be prescriptive of state or district policymaking in regard to measurement choice. Although many districts adopting comprehensive teacher evaluation systems use value-added estimation and external observation ratings, relatively few use student survey responses. Instead, districts may include other teacher performance measures (e.g., measures of teacher professionalism, principal ratings, school value-added), none of which are investigated in the MET studies. The extent to which our simulation's results will generalize to these alternative measures depends on their correlation with other included measures and their correlation with LTVA. If these alternative measures show small, positive correlations with value-added (consistent with the findings of the MET studies' measures), they will presumably behave similarly when adopted into a district's evaluation system.

Second, we wish to clarify that our use of the 2010 and 2012 MET studies' correlation estimates implicitly adopts that project's approach to measuring teacher quality—namely, that underlying teacher quality is best represented by the LTVA productivity over a teachers' career and that the assumptions the project uses to derive these correlations with LTVA are valid. This focus is, admittedly, a narrow view of teacher quality; other elements of teacher performance and professionalism may be valuable to students and the schools that employ them but may be poorly correlated with value-added estimates. Perhaps different performance measures may be picking up different factors of teacher quality, and focusing on LTVA implicitly ignores these other dimensions. We concede that teacher quality could be defined more broadly but believe this approach is still useful as it does capture important differences in teachers that are correlated with long-term student outcomes of interest (e.g., Chetty, Friedman, & Rockoff, 2011). We wish to highlight, though, that our primary objective is understanding how districts' approaches for combining multiple measures impact the identification of teacher quality on some target criterion; the exact target criterion could be changed and is not the primary objective of our investigation. Rather, districts may use an alternative target criterion, which may be optimizing the expected performance of the workforce jointly measured along several dimensions (as explored in

Mihaly et al., 2013, using the MET study data). So long as the target criterion can be quantitatively defined using performance measures and relies on multiple measures for inference, the lessons from our analysis will still apply.

Third, attempting to classify teachers into discrete performance categories, as districts do in these evaluation systems, introduces the possibility of misclassifying teachers into these categories. As described earlier, misclassification rates are key metrics that we use to compare the goodness of fit between these approaches, but we emphasize that these misclassification rates are specific to the framework we establish for this simulation here and are not readily comparable to different frameworks. For instance, consider a different evaluation system that attempts to classify teachers into one of 100 effectiveness categories according to their percentile ranking on the target criterion. Under this system, because the target classifications are so fine grained and so numerous, we would expect very few teachers in such a system to be correctly classified; however, because it provides such rich information about variation in observed performance across the workforce, this system could still be very useful in spite of high misclassification rates. Alternatively, documented evaluation systems currently in place in some districts apparently assign teachers into two categories, "unsatisfactory" and "satisfactory," corresponding to roughly the bottom 1 percent and top 99 percent of teachers, respectively (Weisberg et al., 2009). Even though the technical misclassification error of this kind of system will be extremely low, such systems are decidedly unhelpful in differentiating teachers who are known to show large degrees of variation in productivity. As a general rule, categories intended to capture small segments of the workforce (such as the ME and IE ratings used here) will have higher misclassification rates, while those intended to capture larger swaths of the workforce (the E rating) will have lower misclassification rates, all things equal. Thus, misclassification rates are a function of both the evaluation system's inference and prediction process and the properties of the classification framework onto which the evaluation system is assigning teachers.

This situation raises the issue of what the "optimal" teacher classification framework looks like and how evaluation systems should attempt to map onto it. At a minimum, identifying high and low performers would require at least two different performance categories, although there is no clear consensus on how many different categories of teacher quality should be differentiated, nor is there consensus on how much of the teacher workforce should fall into these different categories. As a practical matter, the design of these classifications should probably reflect the consequences attached to each category: If districts treat teachers with exceptionally low performance (the IE rating) the same as those with slightly higher performance (the ME rating), there is no value in making this distinction. Yet, this distinction does not imply that districts are responding appropriately to meaningfully different levels of performance, and some researchers have argued that districts not only should differentiate the workforce more but also should treat them differentially (Weisberg et al., 2009). Many of the districts that have recently reformed their teacher evaluation systems have implemented a system similar to the framework we adopt for our simulation here, so our results should be broadly generalizable. A broader discussion on the underlying classification of teachers on these performance dimensions is beyond the scope of this analysis.

Results

We conduct the simulation with 1,000 iterations as described earlier and compile the findings. Table 1 reports these baseline results of our simulation. Each column represents a different approach to combining multiple measures: column 1 reports the error-minimizing approach, column 2 reports the numeric approach, column 3 the hybrid approach, and column 4 the profile approach. The fit statistics for each approach, which were described in the Simulation section, are reported in the rows, along with the 95 percent confidence interval for the statistic based on the 1,000 iterations (reported in the

brackets).[15] These confidence intervals represent the likely range of variation that the values of these statistics will take on in a workforce with the given sample size of 500 teachers; in practice, larger (smaller) samples will reduce (increase) the span of these confidence intervals.

Table 1. Simulation Results on Goodness of Fit Against True Classifications

		Error-Minimizing Approach	Numeric Approach	Hybrid Approach	Profile Approach
Ineffective (IE)	Correct	0.048	0.046	0.015	0.007
(actual share = 10%)	Misclassified	0.052	0.054	0.010	0.004
Minimally Effective (ME)	Correct	0.023	0.021	0.037	0.030
(actual share = 10%)	Misclassified	0.078	0.079	0.147	0.105
Effective (E)	Correct	0.433	0.427	0.419	0.404
(actual share = 60%)	Misclassified	0.167	0.173	0.177	0.172
Highly Effective (HE)	Correct	0.115	0.112	0.108	0.129
(actual share = 20%)	Misclassified	0.085	0.089	0.087	0.150
Overall percentage correctly identified		0.619	0.606	0.579	0.570
		[0.5800, 0.6560]	[0.5640, 0.6440]	[0.5400, 0.6180]	[0.5300, 0.6100]
Percentage correctly classified, given identification in tails		0.464	0.447	0.397	0.392
		[0.4100, 0.5150]	[0.3900, 0.5000]	[0.3436, 0.4528]	[0.3442, 0.4423]
Extreme error rate		0.003	0.004	0.007	0.009
		[0.0000, 0.0080]	[0.0000, 0.0100]	[0.0000, 0.0160]	[0.0020, 0.0180]
Average long-term value-added for E or HE teachers		0.251	0.242	0.233	0.185
		[0.1586, 0.3467]	[0.1523, 0.3365]	[0.1393, 0.3247]	[0.0928, 0.2746]
Difference in long-term value-added between tails		2.004	1.934	1.839	1.859
		[1.7703, 2.2273]	[1.6995, 2.1629]	[1.5968, 2.0700]	[1.6128, 2.1030]
Ratio of overstating to understating effectiveness		1.004	1.004	1.187	2.415
		[0.9321, 1.0761]	[0.9358, 1.0737]	[1.0636, 1.3258]	[2.0650, 2.8367]

Note: Simulation results based on 1,000 iterations, using a sample size of 500 teachers. Cells report the mean fit statistic across 1,000 iterations and report the simulation-based 95 percent confidence intervals in brackets. By construction, ineffective teachers constitute the bottom 10 percent of the simulated sample on the underlying teacher quality dimension; minimally effective teachers constitute those in the 11th to 20th percentile; effective teachers are the middle 60 percent of the distribution; and highly effective teachers are those in the top 20 percent of the distribution. Decision rules for categorizing teacher performance under each of the four models are described in the text, as well as the interpretation of the summary statistics at the lower panel of the table.

All of the models presented here, even the error-minimizing approach, demonstrate only modest levels of accuracy, with the overall percentage of teachers correctly categorized ranging from 57 percent to 62 percent. Inspection of the correct and misclassified rates by identified category in the upper panel of Table 1 shows that the likelihood of misclassification is negatively related to the size of

[15] The confidence intervals are determined by referencing the 2.5th and 97.5th percentiles of the particular fit statistic's distribution over the 1,000 iterations of the simulation.

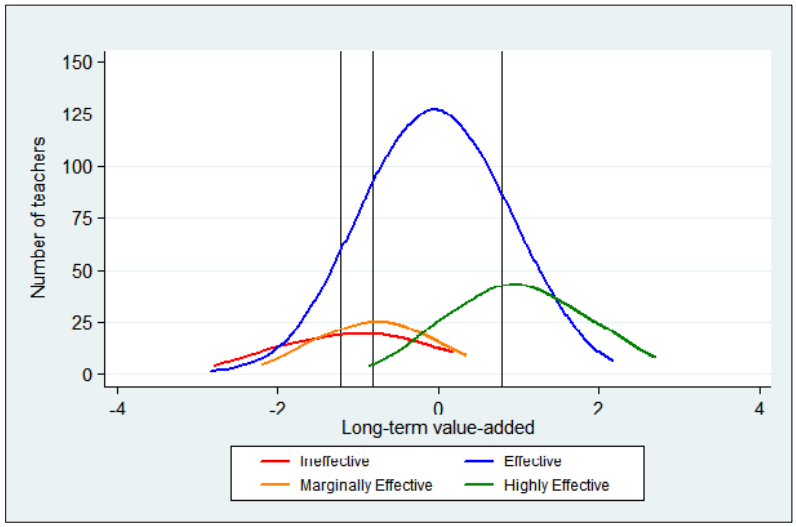
the category into which teachers are binned (e.g., misclassification is highest in the ME category and lowest in the E category), as discussed previously. This result is true for all four of the approaches presented here. Hence, we also report the percentage correctly classified for those teachers identified in the tails (IE, ME, or HE), and these percentages are considerably lower, ranging from 39 percent to 46 percent. Across all approaches, the likelihood of an extreme error (i.e., identifying a teacher as HE if his or her true performance was ME or lower, or vice versa) is less than 1 percent.

Comparing these classification statistics across the four columns, we see that the error-minimizing approach performs the best of all options, followed by the numerical, hybrid, and profile approaches, in that order. This ordering is in line with our predictions based on the progressively larger prediction error associated with these approaches. The differences between the approaches in columns 1 and 4 (the error-minimizing approach and the profile approach) are significant at the 95 percent level; yet, these differences across the three approaches commonly used in districts are generally small and are not significantly different.

Figure 3 graphically depicts the distributions of the underlying LTVA for teachers in each of the identified classifications from a single iteration of simulated data (this figure uses the error-minimizing approach to identify teacher effectiveness ratings). The three vertical reference lines represent the actual threshold between the effectiveness categories on the target criterion. This graphic provides a visual image of the misclassification inherent in this process of combining multiple measures to infer performance. For example, the teachers to the left of the very first reference line are those who are truly IE teachers, but a relatively small fraction of these teachers are identified as IE, a slightly smaller fraction are identified as ME, and a large fraction of these true IE teachers are identified as E. From this graphic, it is clear that although there are meaningful differences in LTVA in each of the four identified categories, large

amounts of overlap exist among these models where teachers are misclassified because of measurement error.

Figure 3. Distribution of Long-Term Value-Added by Identified Effectiveness Rating



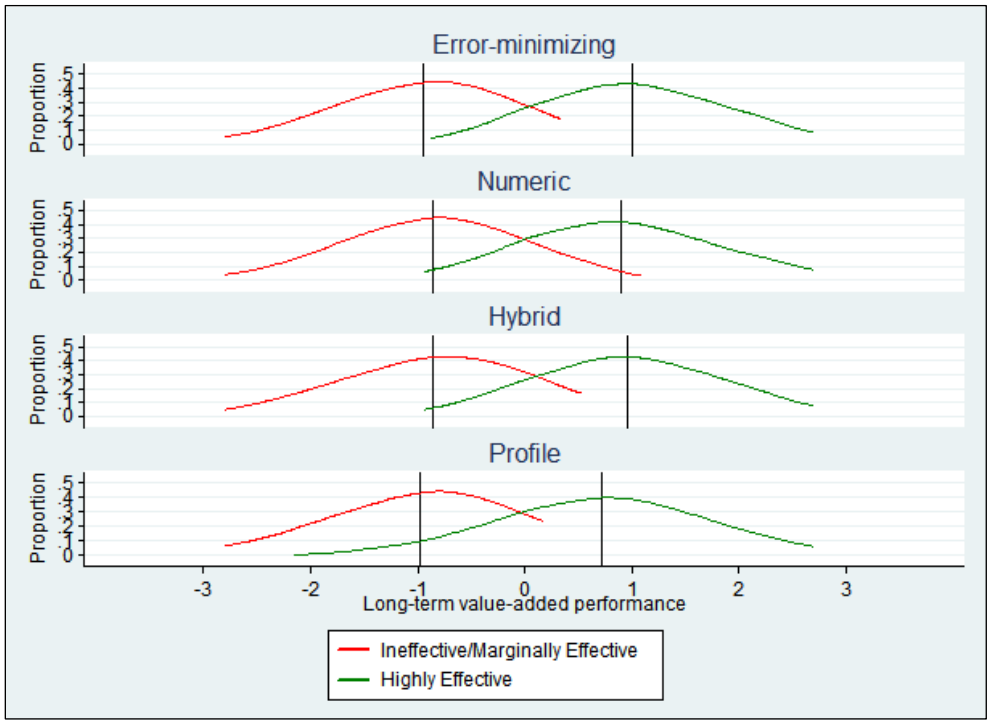
Referring back to Table 1 (see page 15), the next two statistics deal with the productivity of the teacher workforce identified under these approaches, using LTVA. The first statistic reports the average LTVA of the teachers identified as E or better; the evaluation system implicitly endorses these teachers. This statistic shows these teachers combined are 0.19 to 0.25 standard deviations of LTVA better than the population mean, depending on the approach, though none of these estimates are statistically different from each other.

Next, the differences in LTVA between the tails of the distribution vary across these approaches; but, again, none of the approaches are significantly different from the others. Figure 4 provides a representation of this statistic, using a single iteration of the simulated data. As shown, the error-minimizing approach shows the largest gap between the teachers in the upper and the lower tails (i.e., discriminates on LTVA more effectively), with the profile approach showing the smallest gap. Also note how the distributions progressively overlap as the gap decreases in magnitude;

the profile approach in particular shows a surprising level of commonality between the teachers

included in these categories intended to identify teachers at the extremes of the distribution of LTVA.

Figure 4. Measuring the Difference in Long-Term Value-Added Between the Tails



The final row of Table 1 (see page 15) indicates the ratio of overstating to understating effectiveness, which we interpret as a measure of bias in the aggregation approach. Results near 1 indicate no bias, while those greater than 1 indicate a bias that disproportionately benefits teachers. No evidence of bias appears in the error-minimizing and numeric approaches, where the likelihood of overstating a teacher's performance is proportional to the likelihood of understating a teacher's performance. A mild bias is present in the hybrid approach, where teachers are 19 percent more likely to have their performance overstated, and a strong bias is present in the profile approach, where teachers are more than twice as likely to have their true LTVA overstated (relative to their likelihood of being understated). This bias also is evident in inspection of the correct and misclassified percentages by effectiveness rating in the table's upper panel. Note that the profile approach identifies only 1 percent

of teachers as IE (where the actual share is 10 percent) while identifying 28 percent of teachers as E (actual share: 20 percent). These bias results are in line with our expectations, and the differences in these statistics are statistically significant from the value of 1 for the latter two approaches.

Summarizing the results from Table 1, we see evidence consistent with our expectations about the relative performance of these models. Namely, of the three commonly used approaches, the numeric approach is generally preferred as it has lower misclassification rates and identifies higher levels of LTVA and larger gaps between the tails without introducing any bias. Yet, with the exception of bias, these advantages attributed to the numeric approach are not statistically different from the other approaches that introduce greater error into the problem. As far as bias is concerned, the presence of this bias may be acceptable to some districts because it overstates true teacher effectiveness (hence, there likely will be little resistance to this approach from teachers). However, we wish to clarify that a bias favoring teachers implicitly penalizes students to whom these marginal teachers are assigned. Hence, it is incorrect to consider the bias entirely innocuous: It may simply be more difficult to determine its effect on student outcomes over time.

Adjusting Component Weights

We next investigate how our comparisons of these various approaches may change using a different set of weights on the three component measures that are combined to infer a teacher's LTVA. Up to this point, we have used a baseline weighting scheme allotting 50 percent of the total combined score to value-added estimates, 35 percent to observation ratings, and 15 percent to student survey results; this weighting is reasonably aligned with the relative reliabilities of these measures against LTVA (see discussion in footnote 12). Yet, when districts choose component weights through a political or bargaining process, these chosen weights may be very different from the optimal weights that minimize

the mean squared error on the district's prediction of teachers' performance on the target criterion.[16] Recall that the only difference between the error-minimizing and numeric approaches is the use of different component weights, and this approach is the source of prediction error under the numeric approach. If districts use weights that are near the optimal weights, the error introduced by using separate weights is small; if districts use weights that are far from the optimal weights, the error will increase accordingly. Hence, we know that using alternate weights further removed from the optimal weights will reduce the efficiency of the three commonly used approaches overall, although we are unclear on whether the differences between the approaches will be statistically significant. To investigate the influence of weights on the performance of these three approaches, we employ alternate weights that give 25 percent to value-added estimates, 50 percent to observation ratings, and the remaining 25 percent to student surveys.

Table 2 presents the results using this alternate weighting scheme. Note that the error-minimizing approach in this table is equivalent to the results reported in Table 1, because the error-minimizing approach uses the same optimal weights in both tables. A similar ordering to Table 1 is generally apparent between the three commonly used approaches, where the numeric approach shows the best fit statistics and the profile approach shows the worst. Under these alternate weights, however, the differences between the profile approach and the other two are more pronounced and are statistically significant in the case of the overall percentage correct, the percentage correct in the tails, and the LTVA gap between the tails. The differences between the numeric and hybrid approaches are not statistically significant, with the exception of the bias measure. Interestingly, our bias statistic does not follow these general ordering patterns; the bias favoring teachers is apparently largest under the hybrid approach

[16] The recent analysis from Mihaly et al. (2013) concludes that using roughly equally balanced weights can be optimal for districts if the target criterion of the evaluation system is one that maximizes the expected performance of the workforce on all three performance measures and if it takes the underlying correlation structure between the measures as given. This result does not imply that districts choosing to equally weight their measures will be doing so optimally. If these districts have a different target criterion or use measures that vary in their correlation structure against the target criterion, the optimal weighting will shift.

(this is statistically significant and greater than the numeric approach). There is no detectable bias in the profile approach, contrary to our expectations; this result may be because of the level of prediction error swamping the effect of the underlying bias.

Table 2. Simulation Results on Modifying Parameters

Using an alternate set of weights that down-weight value-added estimates

		Error-Minimizing Approach	Numeric Approach	Hybrid Approach	Profile Approach
Ineffective (IE)	Correct	0.048	0.039	0.008	0.009
(actual share = 10%)	Misclassified	0.052	0.061	0.005	0.007
Minimally Effective (ME)	Correct	0.023	0.018	0.029	0.035
(actual share = 10%)	Misclassified	0.078	0.082	0.130	0.174
Effective (E)	Correct	0.433	0.407	0.412	0.384
(actual share = 60%)	Misclassified	0.167	0.193	0.210	0.226
Highly Effective (HE)	Correct	0.115	0.099	0.093	0.066
(actual share = 20%)	Misclassified	0.085	0.101	0.113	0.101
Overall percentage correctly identified		0.619	0.563	0.542	0.493
		[0.5800, 0.6560]	[0.5260, 0.6020]	[0.5040, 0.5800]	[0.4520, 0.5340]
Percentage correctly classified, given identification in tails		0.464	0.391	0.345	0.279
		[0.4100, 0.5150]	[0.3400, 0.4450]	[0.2874, 0.4022]	[0.2268, 0.3350]
Extreme error rate		0.003	0.010	0.014	0.019
		[0.0000, 0.0080]	[0.0020, 0.0180]	[0.0060, 0.0240]	[0.0080, 0.0320]
Average long-term value-added for E or HE teachers		0.251	0.209	0.165	0.185
		[0.1586, 0.3467]	[0.1172, 0.2977]	[0.0741, 0.2549]	[0.0925, 0.2834]
Difference in long-term value-added between tails		2.004	1.670	1.534	1.243
		[1.7703, 2.2273]	[1.4151, 1.9158]	[1.2793, 1.7983]	[0.9751, 1.4931]
Ratio of overstating to understating effectiveness		1.001	1.000	1.203	0.941
		[0.9305, 1.0721]	[0.9270, 1.0782]	[1.0870, 1.3333]	[0.8542, 1.0232]

Note: Simulation results based on 1,000 iterations, using a sample size of 500 teachers. Cells report the mean fit statistic across 1,000 iterations and report the simulation-based 95 percent confidence intervals in brackets. By construction, ineffective teachers constitute the bottom 10 percent of the simulated sample on the underlying teacher quality dimension; minimally effective teachers constitute those in the 11th to 20th percentile; effective teachers are the middle 60 percent of the distribution; and highly effective teachers are those in the top 20 percent of the distribution. Decision rules for categorizing teacher performance under each of the four models are described in the text, as well as the interpretation of the summary statistics at the lower panel of the table.

In summary of these results, we find evidence that the choice of aggregation approach becomes more consequential when using weights that increase the distance from the optimal, empirically determined weights. Using these alternate weights affects our results in two important ways. First, as expected, we found a general decline in the efficiency of all three competing approaches to discriminate

on LTVA with corresponding increases in misclassification. Second, the differences in the fit statistics between the approaches grew; and, in this case, the profile model was significantly inferior on several statistics. The numeric and hybrid models were still close enough to be statistically indistinguishable, with the exception of bias.

Using Reliability-Adjusted Performance Measures

We also investigate how the use of reliability-adjusted performance measures affects the identification of teachers in these simulated districts. Given the measurement error associated with any of the measures used in the aggregation process (both the component measures themselves and the combined score), districts may choose to scale performance measures according to the reliability of the given measure against the target criterion before discretely categorizing them. Glazerman et al. (2011) encourage the use of this adjustment technique as a means to reduce error in misclassifying high- or low-performing teachers. By design, this adjustment method will identify far fewer teachers in the tails of the distribution (IE, ME, and HE ratings) because their performance is not statistically distinguishable from mean performance. However, we are unsure how this approach will affect misclassification rates, differences in LTVA, or bias across the models.

We perform this adjustment for each of the four approaches using the baseline weighting scheme and present these results in Table 3.[17] Looking at the upper panel reporting the classification percentages by effectiveness rating, there are indeed far fewer teachers in the tails; the hybrid and profile approaches identify no teachers as IE, and fewer than 4 percent are identified under the error-minimizing and numeric approaches. The ME and HE ratings likewise identify many fewer teachers than the actual share in these

[17] As described in the Appendix of Glazerman et al. (2011), this technique is performed by first estimating the reliability of the given performance measure that is to be categorized by regressing it alone against the target criterion. (The authors use student test scores in the following year as the target criterion; we use simulated LTVA.) The coefficient from this regression is then multiplied with the performance measure to be categorized (scaling it by the reliability) before determining where it falls in comparison to the predetermined threshold performance levels. Under the error-minimizing and numeric approaches, we perform this adjustment on the combined score just before teachers are categorized into their summative measures; the hybrid and profile approaches perform this adjustment on the individual performance measures themselves before their initial categorization.

classifications. The overall percentage correct and percentage correct in the tails are significantly higher across the board, and the extreme error rate is much lower using this reliability-adjusted method. As a consequence of disproportionately identifying teachers as E, the average LTVA of the E or better teachers is significantly lower across approaches than what was observed in Table 1. On this measure, the numeric approach is not statistically different from the error-minimizing approach but is significantly greater than both the hybrid and profile models, both of which contain 0 in their confidence intervals. Because the teachers with exceptionally low or high performance are the ones identified in the tails (i.e., the tails are more selective), these LTVA gap measures are considerably larger than what were estimated in Table 1. The bias toward teachers is again present in both the hybrid and profile approaches, while the error-minimizing and numeric approaches show no evidence of a bias.

Table 3. Simulation Results on Modifying Parameters

Using reliability-adjusted approach when categorizing teacher performance

		Error-Minimizing Approach	Numeric Approach	Hybrid Approach	Profile Approach
Ineffective (IE)	Correct	0.024	0.020	0.000	0.000
(actual share = 10%)	Misclassified	0.013	0.012	0.000	0.000
Minimally Effective (ME)	Correct	0.020	0.019	0.007	0.007
(actual share = 10%)	Misclassified	0.063	0.061	0.030	0.028
Effective (E)	Correct	0.521	0.525	0.555	0.553
(actual share = 60%)	Misclassified	0.238	0.252	0.295	0.295
Highly Effective (HE)	Correct	0.081	0.074	0.074	0.076
(actual share = 20%)	Misclassified	0.040	0.038	0.038	0.041
Overall percentage correctly identified		0.646	0.638	0.637	0.636
		[0.6120, 0.6780]	[0.6040, 0.6680]	[0.6100, 0.6640]	[0.6100, 0.6640]
Percentage correctly classified, given identification in tails		0.518	0.505	0.547	0.550
		[0.4386, 0.5963]	[0.4153, 0.5840]	[0.4483, 0.6595]	[0.4487, 0.6575]
Extreme error rate		0.001	0.001	0.001	0.001
		[0.0000, 0.0040]	[0.0000, 0.0040]	[0.0000, 0.0020]	[0.0000, 0.0040]
Average long-term value-added for E or HE teachers		0.164	0.149	0.060	0.056
		[0.0835, 0.2434]	[0.0634, 0.2354]	[-0.0240, 0.1497]	[-0.0273, 0.1456]
Difference in long-term value-added between tails		2.388	2.357	2.691	2.699
		[2.1852, 2.5868]	[2.1426, 2.5594]	[2.3226, 3.0317]	[2.3254, 3.0588]
Ratio of overstating to understating effectiveness		1.154	1.142	1.746	1.819
		[0.8248, 1.5873]	[0.8901, 1.4458]	[1.4384, 2.2104]	[1.4745, 2.3208]

Note: Simulation results based on 1,000 iterations, using a sample size of 500 teachers. Cells report the mean fit statistic across 1,000 iterations and report the simulation-based 95 percent confidence intervals in brackets. By construction, ineffective teachers constitute the bottom 10 percent of the simulated sample on the underlying teacher quality dimension; minimally effective teachers constitute those in the 11th to 20th percentile; effective teachers are the middle 60 percent of the distribution; and highly effective teachers are those in the top 20 percent of the distribution. Decision rules for categorizing teacher performance under each of the four models are described in the text, as well as the interpretation of the summary statistics at the lower panel of the table.

Overall, these results using reliability-adjusted performance measures improve the correct classification rates for all teachers, and the LTVA gaps between the tails are much larger relative to the results without these adjustments. Although these are improvements across the board on these measures, there are important declines on the average LTVA among those identified as E or better, and the hybrid and profile models perform particularly poorly on this measure. Thus, these improvements in correct classification among teachers come at the expense of lower efficiency in identifying LTVA for most teachers in the workforce. For any districts, however, that choose to trade the reduction in overall efficiency for the increased confidence of identifying teachers at the extremes of the distribution, the

numeric approach must be the preferred approach to use because the hybrid and profile approaches are far inferior when using reliability-adjusted measures.

Conclusion and Discussion

The title of this paper asks whether common approaches to combining multiple performance measures undermine districts' personnel evaluation systems. Based on the results of our simulations presented in this paper, we conclude yes, these approaches can undermine the evaluation system's objectives in some contexts. The three prototypical approaches that districts commonly use, and that we investigate here, add error and in some cases bias into the district's ability to infer a teacher's true performance on the target criterion. Depending on the way these performance measures are combined to categorize teacher performance, the additional error and bias can be large enough to render the evaluation system almost useless. Specifically, we find three primary results.

First, among the three commonly used approaches in practice, the numeric approach showed the best fit across all of the different scenarios we investigated. We expected this result, given that this approach adds the least amount of error into the district's prediction on the target criterion and does not introduce any bias. Under both the baseline simulation and simulation using reliability-adjusted performance measures, the differences between the optimal error-minimizing approach and the numeric approach used in practice were not statistically significant. Thus, in these circumstances, the numeric approach can approximate the optimal outcome in a district. When using component weights that were further misaligned with the underlying reliabilities, the differences between the error-minimizing and numeric approaches were significant but were still smallest compared with the other commonly used approaches.

Second, the hybrid and profile approaches, also commonly used in practice, in some circumstances perform close enough to the numeric model that the additional error and bias introduced under these

models may be acceptable to some districts. Under the baseline scenario, misclassification rates and the LTVA fit statistics of both approaches were not statistically distinguishable from the numeric model; again, the hybrid model was not statistically distinguishable from the numeric model when component weights were less well aligned with the target criterion. Both models, however, introduce a bias that favors teachers (i.e., these models are more likely to overstate rather than understate teachers' true performance), and this bias is generally statistically significant in our simulations. Yet, we emphasize that this bias against teachers implicitly works against students, as overstating teacher performance allows low-performing teachers to continue teaching without consequence. And finally, when using reliability-adjusted performance measures, both the numeric and hybrid models identify very few teachers as IE or ME; thus, the average LTVA of those identified E or higher was not statistically different from zero. In other words, the evaluation system could be entirely ineffective in raising the LTVA of the workforce under these circumstances.

And third, using reliability-adjusted performance measures to categorize teachers' performance generally improves correct classification rates overall and is attended with more selective identification of high and low performers. Yet, this method obtains these improved classification rates by disproportionately identifying teachers as effective, thereby reducing the district's ability to identify high or low performance. If districts are willing to trade the lower effectiveness of the system overall for the increased precision in the tails, the numeric model shows superior performance under this method.

This investigation has some noteworthy limitations on how these results may be interpreted. First, we adopt the intercorrelations between performance measures from the MET studies at face value, although these estimates are rigorously validated. In practice, districts may not understand the reliabilities between measures or against the target criterion, and in this case,
we may reasonably expect the evaluation system to be less effective overall (i.e., higher misclassification

rates, lower ability to select high and low performers). Second, the approaches we investigate here are simplified abstractions of those that are used in practice; thus, implemented approaches may have some features that carry larger consequences on their ability to adequately discriminate on teacher quality. And finally, we consider only teachers for whom all three simulated performance measures are available, and we do not address how teachers without value-added measures should be evaluated; this distinction is a key issue in implementing comprehensive evaluation systems, but it is beyond the scope of this particular study.

Recommendations

Based on these simulation results, we offer the following recommendations for states and districts constructing their own evaluation systems.

- **Recommendation 1.** Among the three commonly used district approaches investigated here, the numeric approach is the overall preferred approach. Under some circumstances, the hybrid and profile approaches can possibly be used in place of the numeric approach without introducing any statistically significant differences in the fit, although the choice in these circumstances must be carefully weighed because both approaches still introduce a bias favoring teachers.
- **Recommendation 2.** States and districts should clearly articulate the evaluation system's target criterion and understand how the adopted performance measures predict that target criterion. We use LTVA as our target criterion, although districts may pursue their own objectives. Regardless of that choice, haphazardly weighting component measures without understanding the empirically optimal weighting structure may unduly hinder the effectiveness of the evaluation system.

This simulation study has provided insight into the consequences of how performance measures are combined and categorized in identifying high- and low-performing teachers in practice. We further recommend that districts and states conduct their own simulations. Such simulations are a low-cost tool that can help inform how a district's particular design choices may impact the teacher workforce in their schools. By understanding how best to use these performance measures, states and districts can be enabled to implement efficient workforce management policies that promote both quality teaching and student learning in the years to come.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., ... Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute. Retrieved from http://www.epi.org/page/-/pdf/bp278.pdf

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project* (MET Project Research Paper). Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/Preliminary_Finding-Policy_Brief.pdf

Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching* (MET Project Research Paper). Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study* (MET Project Policy and Practitioner Brief). Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Branch, G., Hanushek, E., & Rivkin, S. (2009). *Estimating principal effectiveness* (CALDER Working Paper 32). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.urban.org/uploadedpdf/1001439-Estimating-Principal-Effectiveness.pdf

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (NBER Working Paper 17699). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17699

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: Evaluating evaluation systems.* Washington, DC: Brookings Institution, Brown Center on Education Policy. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2011/4/26%20evaluating%20teachers/0426_evaluating_teachers.pdf

Goldhaber, D., & Chaplin, D. (2012). *Assessing the "Rothstein test": Does it really show value-added models are biased?* (CALDER Working Paper 71). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.caldercenter.org/upload/Assessing-the-Rothstein-Test_wp71.pdf

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *The American Economic Review, 100*(2), 250–255.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the stability of estimated teacher performance. *Economica, 80*(319), 589–-612.

Grossman, P. L., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J. H., Boyd, D. J., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.eric.ed.gov/PDFS/ED511765.pdf

Hanushek, E. A. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 165–180). Washington, DC: Urban Institute Press.

Hanushek, E. A., & Rivkin (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101–136.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy, 6*(1), 18–42.

Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis, 34*(1), 109–121.

Leo, S. F., & Lachlan-Haché, L. (2012). *Creating summative educator effectiveness scores: Approaches to combining multiple measures.* Washington, DC: American Institutes for Research. Retrieved from http://educatortalent.org/inc/docs/Creating%20Summative%20EE%20Scores_FINAL.PDF

McCaffrey, D. F., Sass, T.R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572–606.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching* (MET Project Research Paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

National Council on Teacher Quality (2012). *State of the states 2012: Teacher effectiveness policies.* Washington, DC: Author. Retrieved from http://www.nctq.org/p/publications/docs/Updated_NCTQ_State%20of%20the%20States%202012_Teacher%20Effectiveness%20Policies.pdf

Public Impact. (n.d.). *Building an opportunity culture for America's teachers: Extending the reach of excellent teachers* [Website]. Retrieved from http://opportunityculture.org/

Rockoff, J., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *The American Economic Review*, *100*(2), 261–266.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1), 175–214.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *The Journal of Economic Perspectives*, *24*(3), 97–117.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project. Retrieved from http://widgeteffect.org/downloads/TheWidgetEffect.pdf

Appendix

I. Correlation Matrix

The baseline results used the following correlation matrix to simulate the teacher performance measures. This matrix is based on the largest estimated correlation values reported across the various measures investigated in the 2010 and 2012 MET studies:

Component	Correlation Values			
Underlying teacher quality	1.00			
Value-added estimates	0.69	1.00		
Observation scores	0.34	0.27	1.00	
Student survey scores	0.37	0.43	0.15[a]	1.00

[a] The correlations between observation scores and student surveys were not reported in the MET studies. We inserted the value 0.15, which we presumed to be a conservative estimate.

II. Estimated Coefficients Under the Error-Minimizing Approach

To implement the error-minimizing approach, we regressed the simulated long-term value-added performance values for teachers on the three component measures. The estimated coefficients and the simulation-based 95 percent confidence intervals (based on 1,000 iterations) are presented below. Note that the estimated coefficient is largest on the value-added estimate, next largest for the observation rating, and smallest for the student survey results; the relative weightings are relatively aligned with the baseline component weights used in the simulation though the actual weights vary.

Component	Estimated Coefficient*
Value-added estimates	0.6133 [0.5382, 0.6846]
Observation scores	0.1632 [0.1010, 0.2289]
Student survey scores	0.0818 [0.0129, 0.1508]

*95-percent interval for the estimated coefficient values presented in brackets.