A Composite Estimator of Effective Teaching

Kata Mihaly Daniel McCaffrey Douglas O. Staiger J. R. Lockwood

September 2, 2012

Abstract

This paper explores ways to combine multiple measures of teaching to obtain optimal predictions of target criteria representing valued outcomes. We show that the first steps in combining multiple measures of teaching requires value judgments from policymakers and stakeholders. Once the objective of the composite is established, we show how to optimally combine multiple measures taking correlated measurement error for each measure into account. Finally, we derive the optimal weights as a function of composite component reliability and the correlation of the target criteria, and discuss the implications of changes in reliability induced by data collection strategies on the optimal weights. An upcoming version of this paper will examine how the available data and measurement error in observed measures under various data collection strategies affect the ability to predict the target criteria using data from the Measures of Effective Teaching project.

PRELIMINARY AND INCOMPLETE

1 Introduction

Over the last decade, federal, state and local policy makers and educators have become increasingly concerned with the quality of teaching provided by public school teachers in the United States. In response to the shortcomings of existing evaluation systems and reflecting a shift in focus from teacher qualification to teacher effectiveness, state and large school districts are rapidly developing and adopting new teacher evaluation plans. According to a report from the National Council of Teacher Quality examining teacher evaluation and effectiveness policies, 32 states and the District of Columbia have made changes to state teacher evaluation policy as of October 2011 (NCTQ Report (2011)).

Many of these states require objective evidence on student learning, such as student achievement growth or value-added (VA), to be a significant part of the new evaluation system. However, a large majority of states also require additional measures of teaching to be taken into account. These additional metrics almost universally include observations of teaching and in some cases also include other sources of information such as teacher reflections, student learning objectives, or student survey responses about their classroom experiences.

While the details vary state to state and are not always clearly specified, nearly all of the new policies imply that multiple measures of teaching will be combined to produce a single composite measure of teacher effectiveness. These composite measures are motivated by the need to evaluate teachers for the purpose of professional development, tenure, compensations, and retention decisions.

Although states are developing complex rules for combining multiple measures into a single composite score, there is limited empirical research to guide policymakers on how best to combine measures to achieve specific goals. This paper explores ways to combine multiple measures of teaching to obtain optimal predictions of target criteria representing valued outcomes. We show that the first steps in combining multiple measures of teaching requires value judgments from policymakers and stakeholders. Once the objective of the composite is established, we show how to optimally combine multiple measures taking correlated measurement error for each measure into account. Finally, we derive the optimal weights as a function of composite component reliability and the correlation of the target criteria, and discuss the implications of changes in reliability induced by data collection strategies on the optimal weights. An upcoming version of this paper will examine how the available

data and measurement error in observed measures under various data collection strategies affect the correlation of the composite with target criteria using data from the Measures of Effective Teaching project.

2 Background and Conceptual Framework

Composite indicators have been used widely to evaluate the quality of performance in a number of areas including health care (Jacobs, Smith and Goddard (2004), Reeves et al. (2007), Dimick et al. (2009)), university performance (Johnes (1992), Murias, de Miguel, and Rodríguez (2008), Editor (2008)), water quality (Carr and Rickwood (2008)), managerial performance (Holgstrom and Milgrom (1991)), as well as local and national governments (Freudenberg (2003), Kaufman, Kray, and Mastruzzi (2010), Klugman et al. (2011)). The wide use of composite indicators in these fields is understandable, considering the number of benefits from using a single index to evaluate performance as cited by Saisana and Tarantola (2002) and Mehrens (1990): they can summarize multi-dimensional indicators into a single number which is required for decision making and potentially do so without losing the underlying information, they are easier to interpret than multiple indicators, they facilitate communication to the public, and they promote accountability.

However, as noted by the Saisana and Tarantola, there are a number of drawbacks to using composites: they may invite misleading and simplistic policy conclusions if they are misinterpreted or poorly constructed, they may be misused to support desired policies if the process of constructing them is not transparent or not based on sound principles, they may lead to inappropriate policy conclusions if the dimensions of performance that are difficult to measure are ignored, and they may disguise serious failings on some dimensions and increase the difficulty of focusing remedial action. In addition, Behn (2003) argues that there are multiple distinct purposes to measuring performance and that different types of performance measures are more or less well-suited for the different purposes. Along this same line, Schmidt and Kaplan (1971) note that the use of composites is motivated by an "economic value-to-the-organization" argument in which the single composite is used to make decisions to optimize the organization's goals, whereas retaining multiple component measures is motivated by behavioral psychological arguments to achieve psychological understanding and direct specific changes much the way the balance score card approach was advocated for business as an

alternative to focusing only on profits or composites meant only to optimize profits.

Despite these shortcomings, and because of their desire to provide measures to support decision making, states are moving forward in requiring composite indicators of teacher quality. Although states have developed complex rules for combining multiple measures into a single composite score, there is tremendous desire for guidance on how to combine measures to achieve specific goals. We identified three classes of questions that arise in creating composites: 1) What do stakeholders and experts value in teaching and how are those characteristics to be measured? 2) What are optimal statistical weights and how are they determined from the data? and 3) What is the value, in terms of increased correlation between observed composite performance measures and the target criterion of interest, from collecting and combining different sources of data and how does this vary for different target criteria that stakeholders and experts might identify. We address each set of questions in turn. In the next section, we discuss the essential role for expert opinion in developing composites. We then discuss the optimal weights and how they might be estimated when the composite is chosen to be a weighted sum of the component measures. In an upcoming version of the paper we will estimate the value of various data collection strategies using data from the Measures of Effective Teaching Project (MET).

2.1 Value Decisions

Mehrens (1990) distinguishes between a "clinical" and "statistical" approach to combining measures to create a composite. With the clinical approach judgment is used to create a rule for combining data. With the statistical approach fixed weights are used to combine the data and statistical analyses can be used to derive the weights. Early literature showed that when there is a measure of a criterion that serves as a target for optimizing the statistical weights, then statistical rules provide greater accuracy (Mehrens (1990), Dawes and Corrigan (1974)).

However, Mehrens (1990) also argues that there are advantages to using professional judgments in the development of composites. We strongly agree with this perspective, but argue that this judgment should be used in determining how to combine the qualities of teaching captured by the empirical measures rather than the empirical measures themselves. We argue that at the outset of the process of forming a composite measures, stakeholders must decide upon the concepts (valued

outcomes) to be measured by the composite, the criteria that will serve as proxy measures for these value outcomes, and the aggregation rules for the criteria, including the value weights of each criterion when the aggregation rule is linearly additive.

As noted OECD handbook on teacher evaluation (Nardo et al. (2008)), the first step to creating a composite measure involves defining the underlying concept to be measured by the composite. Expert judgment will be essential for this task. In many applications outside of education, there is a single phenomenon, or valued outcome, that is the intended goal of the composite. For example, the intended goal of a managerial quality index is to capture dimensions of managers that maximize the financial returns to the firm, and indicators of hospital performance can be designed to measure outcomes such as mortality rates. However, the education context is complicated by the fact that it is difficult to define a single valued outcome. Society may value education for multiple reasons, including helping citizens make positive contributions to the economy or achieving personal fulfillment. In addition, there are dimensions of teaching that may be valued for the purpose of professional development and it may be desirable to include these dimensions in the composite measure. Because the teaching is multidimensional with potential multiple valued outcomes, judgments provided by experts and stakeholders must first define the concept to be measured by the composite indicator. They must then decide on the criteria that will be used as proxies for the valued outcomes.

Even if the valued outcome can be agreed upon, stakeholders do not observe those outcomes in a timeframe that is useful for evaluations. In practice, outputs of schooling that are thought to predict valued outcomes can be used as proxy measures. For example, student achievement at the end of the year can be used as an output that predicts college and career readiness. Alternatively, a student's academic perseverance or effort can proxy for likely lifetime outcomes. We can use input measures collected by states such as teacher observations to proxy for other valued outcomes for students and teachers. Importantly, we assume that the proxy variables measure the criteria of interest with measurement error, which has implications for how optimal composites are constructred and their reliability.

Next, stakeholders must decide on a rule for combining the criteria. Three approaches to combining data are the conjunctive, disjunctive, and compensatory models (Gulliksen (1950), Mehrens (1990)). The conjunctive model creates multiple cutoffs on each criterion and gives ratings based

on exceeding these thresholds. For instance, in a simple example, if the experts and stakeholders determine that the aspects of teaching that should be used for decision making are its contributions to a combination of student "learning" and "effort," then a conjunctive model for a composite would classify teaching as exemplary only if it is exemplary for both student learning and effort. The disjunctive model makes classification based on at least one criterion meeting a threshold. For instance, in some states states students can pass a high-school exit exam as long as they pass it on at least one of multiple attempts. The compensatory model allows for high values in one criterion to compensate for low values in other criteria. The canonical compensatory model is the linear additive model which sets the composite equal to a weighted sum of the component criteria. In the context of the compensatory model, experts and stakeholders must also decide on weights that reflect the contribution of each component criterion to their utility. We call these the "value weights" and the resulting weighted sum of the criteria as the "target criterion". In the simple example, where experts and stakeholders focus only on contributions to a combination of student learning and effort, policymakers may decide that learning is a more important component of the composite, and should therefore receive a higher weight.

Note that this process for forming a composite indicator is slightly different than previously described in the literature (Nardo et al. (2008)). Whereas previous work has focused on the weighting and aggregation of proxy measures, we emphasize that with differing and correlated measurement error in the proxy variables, weighting and aggregating the proxies implies value judgments about how the underlying criteria are being combined, and the resulting combinations will likely not reflect the value weights policymakers placed on the underlying criteria.

2.2 Statistical Prediction

Conditional on the value decisions of policymakers it is possible to turn to statistical theory to provide the best estimate of the target criterion as a combination of the observed measures of the individual criteria. Below we will discuss the statistical weights that optimally predict the target criterion. Policy makers must provide the value weights that define the target criterion as a function of the individual criteria.

The available measures of criteria for teachers contain measurement error. For example, using

teacher observations, we would like to capture the teacher's average level of teaching across all days and his or her scores on average across all possible observers. However a teacher's observed scores from classroom observations will vary across days that are observed and the person conducting the rating. Similarly, VA scores will depend in part on the unique achievement scores of the students in the teachers classes and will vary from class to class and year to year around the teachers average VA.

This measurement error contributes to year-to-year instability in measures and could potentially lead to errors in any decisions we make about teachers. The accuracy of a measure of a criterion can be improved by accounting for its measurement error (Morris (1983)). Moreover, it has been shown that we can obtain more accurate estimates of a criterion that is measured with error by predicting it from multiple measures of different criteria provided the criteria are correlated.

3 Optimal Weighting

In general, the best estimate in terms of minimizing the expected squared difference between a target criterion and a composite measure used to estimate it is the expected value, under the statistical model for the data, of the target criterion given the observed data (Lehmann and Casella (1998)). When the data can be modeled a multivariate normal or other elliptical distribution, then the optimal estimate for the target criterion will be a weighted sum of the component measures. This composite will have the highest correlation with the target criterion among all linearly additive (weighted sum) predictors of it. We call these weights the "statistical weights".

To make this more concrete we introduce the following notation and assumptions. We assume that there are K criteria ϕ_{ik}, $k = 1, \ldots, K$, that are of interest on every teacher i. For each criterion we assume there is a measure $Y_{ik} = \phi_{ik} + \epsilon_{ik}$, where ϵ_{ik} are measurement or sampling errors and have mean zero and variance that can vary among teachers. We let $\boldsymbol{\phi}_i = (\phi_{i1}, \ldots, \phi_{iK})'$, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iK})'$, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iK})'$, $\boldsymbol{\mu}$ equal the mean of $\boldsymbol{\phi}_i$ and $\boldsymbol{A}$ and $\boldsymbol{E}_i$ equal the variance-covariance matrices for $\boldsymbol{\phi}_i$ and $\boldsymbol{\epsilon}_i$ respectively. Both the ϕ_{ik} and ϵ_{ik} can be correlated among components for the same teacher so that both $\boldsymbol{A}$ and $\boldsymbol{E}_i$ may have nonzero off-diagonal elements.

We further assume that the target criterion is $\boldsymbol{\lambda}'\boldsymbol{\phi}_i$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)'$ is the set of value

weights as specified by the experts and stakeholders.[1] The value weights may contain zeros. For example, if the measures include measures of teaching inputs such as those captured by classroom observations, some stakeholders might not want to include those in the target criterion because they do not directly measure outcomes. In that case the corresponding λ_k might equal zero. Alternatively, stakeholders might desire an optimal measure of each of multiple target criteria to support the behavioral psychological use of the performance measurement, and each might have separate $\boldsymbol{\lambda}$ with all but one element equal to zero.

We let $\eta_i = \boldsymbol{\lambda}'\boldsymbol{\phi}_i$, and the goal is to find a linear combination or weighted sum of the observed values of a subset of component measures $\boldsymbol{X}_i = \boldsymbol{H}\boldsymbol{Y}_i$ where $\boldsymbol{H}$ contains $\ell \leq K$ rows of a $K \times K$ identity matrix that optimally predicts this target criterion.[2] We allow the component measures used in the composite to be a subset of the potential measures $\boldsymbol{Y}_i$ to support the estimation of the value of various data collection strategies that might not collect all potential measures. As noted above, the optimal predictor of η_i is the conditional mean of η_i given the measures, $E[\,\eta_i \mid \boldsymbol{X}_i\,]$ (Lehmann and Cassella (1998)). If the measures are multivariate Gaussian or otherwise follow an elliptical distribution, then

$$E[\,\eta_i \mid \boldsymbol{X}_i\,] = \boldsymbol{\lambda}^{'}\boldsymbol{\mu} + \boldsymbol{\lambda}^{'}\boldsymbol{A}\boldsymbol{H}^{'}(\boldsymbol{A}^{*} + \boldsymbol{E}_i^{*})^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}*) \tag{1}$$

where $\boldsymbol{A}^{*} = \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}^{'}$, $\boldsymbol{E}_i^{*} = \boldsymbol{H}\boldsymbol{E}_i\boldsymbol{H}^{'}$ and $\boldsymbol{\mu}^{*} = \boldsymbol{H}^{'}\boldsymbol{\mu}$. The covariance between η_i and $\boldsymbol{X}_i$ equals $\boldsymbol{\lambda}^{'}\boldsymbol{A}\boldsymbol{H}^{'}$, so the optimal weights are the population regression coefficients of a regression of η_i on the component measure. If the variance-covariance matrix of the measurement errors varies across teachers the weights will also vary. We cannot directly estimate the coefficients through a regression of η_i on $\boldsymbol{X}_i$ because we do not observe η_i, however, provided we have a measure of each X_{ik} and each Y_{ik} with a nonzero value weight we can develop estimates.[3]

[1] The value weights will need to account for differential scaling of the criteria. Differences in scaling will typically be removed by standardizing the variables to have equal variance of one. If stakeholders specify their weights, $\boldsymbol{\lambda}^{*}$, in terms of standardized variables, then the weights for the unstandardized variables are $\boldsymbol{\lambda} = \boldsymbol{D}^{-1/2}\boldsymbol{\lambda}^{*}$, where $\boldsymbol{D}^{-1/2}$ is a diagonal matrix with elements equal to the reciprocals of the square roots of the diagonal elements of $\boldsymbol{A}$.

[2] $\boldsymbol{H}$ is a selection matrix that selects a subset of the measures of $\boldsymbol{Y}_i$.

[3] The estimation of the components of Equation 1 will be discussed in the next version of this paper.

3.1 Estimation of Optimal Weights

A straightforward method to estimate the coefficients is to obtain two repeated estimates of $\boldsymbol{Y}_i$ for every teacher, call these $\boldsymbol{Y}_{i1}$ and $\boldsymbol{Y}_{i2}$, calculate $z_{i2} = \boldsymbol{\lambda}'\boldsymbol{Y}_{2i}$ and $\boldsymbol{X}_{i1} = \boldsymbol{H}\boldsymbol{Y}_{i1}$, and regress z_{i2} on $\boldsymbol{X}_{i1}$.[4] Assuming measurement errors are independent across the multiple measures and there is no year-to-year or section-to-section variance in errors (which is roughly consistent with the data in our empirical investigations), the covariance between z_{i2} and $\boldsymbol{X}_{i1}$ equals the covariance between η_i and $\boldsymbol{X}_{i1}$. However, if $\boldsymbol{E}_i^*$ are not constant then the regression will not yield consistent estimates of the optimal weights for each teacher. The deviation in the weights from the optimal weights will depend on the variance in the $\boldsymbol{E}_i^*$ across teachers and limited simulations studies suggest that under likely ranges for that variance, the regression estimates will be nearly efficient. We explore this further in our empirical investigations.

An alternative to regression that can yield consistent estimates of the weights is to generate consistent estimates of $\boldsymbol{A}^*$ and $\boldsymbol{E}_i^*$ and plug these into Equation 1. Both maximum likelihood and method of moments are possible for this estimation.

3.2 Correlation with the Target Criterion

A key question in the development of a composite is a determination of the value of different component measures given a target criterion determined to be of interest to stakeholders. We cannot assess the cost of data collection for individual districts, but we can assess the value of data for improving the predictions of the target criterion. We measure the value of data by its contribution to the correlation between the composite and the target criterion. Data can differ depending on the variables measured and the sampling plan for that measurement which will affect the size of the measurement error. For instance, observation scores based on four observations of a teacher will have smaller measurement error than scores based on two observations, and student survey-based measures will have less measurement error if all sections taught be a secondary teacher are surveyed than if only a single section is surveyed.

We also use the correlation between the composite and the target criterion to compare alternative methods for constructing a composite. For example, we can evaluate how adding measures to the

[4]The repeated measures may come from multiple years or multiple sections of data for the same teacher.

composite increases the correlation of the composite with the target criterion, and thereby quantify the value of the additional measures.

We follow Kane and Staiger (2002) in using the correlation between the composite measure and the target criterion as a measure of merit for our estimates. Because this measure is a function of $\boldsymbol{E}_i$, this is a function of i, but in the scenarios we will consider, we use the same value $\boldsymbol{E}$ across all teachers. Note that $\boldsymbol{E}$ is a function of the number of students, sections, raters and videos as specified in the data reliability scenarios. Assuming $\boldsymbol{H}$ contains K rows such that all measures of $\boldsymbol{Y}_i$ are selected, the correlation of the optimally weighted composite with the target criterion η_i is

$$R_{opt} = \sqrt{\frac{\boldsymbol{\lambda}'\boldsymbol{A}(\boldsymbol{A}+\boldsymbol{E})^{-1}\boldsymbol{A}\boldsymbol{\lambda}}{\boldsymbol{\lambda}'\boldsymbol{A}\boldsymbol{\lambda}}} \tag{2}$$

We can also examine the fit statistic for any given composite, $\boldsymbol{\gamma}'\boldsymbol{y}$, where $\boldsymbol{\gamma}$ is the vector of weights for the linear combination and could represent any weighting scheme, such as equal weights or policy-based weights. The fit statistic in this case is defined as

$$R_{\boldsymbol{\gamma}} = \frac{\boldsymbol{\lambda}'\boldsymbol{A}\boldsymbol{\gamma}}{\sqrt{\boldsymbol{\gamma}(\boldsymbol{A}+\boldsymbol{E})\boldsymbol{\gamma}\boldsymbol{\lambda}'\boldsymbol{A}\boldsymbol{\lambda}}} \tag{3}$$

4 Implications

4.1 Optimal Weights

In this section we consider a simple example to examine how the reliability of the components and the correlation in the criteria impacts the optimal weight for each component. Assume there are two components, with corresponding $\boldsymbol{A}$ matrix, with elements ν_1^2 and ν_2^2 on the diagonal, and $\rho\nu_1\nu_2$ on the off diagonal, where ρ is the correlation between criteria 1 and criteria 2. Similarly, the $\boldsymbol{E}$ matrix has elements ω_1^2 and ω_2^2 on the diagonal. In addition, assume that the measurement error between the two components are uncorrelated, such that the off-diagonal element of $\boldsymbol{E}$, $\omega_{12} = 0$.[5] Define the reliability of component 1, $R_1 = \frac{\nu_1^2}{\nu_1^2+\omega_1^2}$, and similarly the reliability of component 2, $R_2 = \frac{\nu_2^2}{\nu_2^2+\omega_2^2}$.

The optimal weights for predicting each of the criteria in this simplified example are given by the

[5] This is a realistic assumption if we assume that one of the criteria is value added, and the other criteria is teacher observations.

2×2 matrix $\boldsymbol{A}(\boldsymbol{A}+\boldsymbol{E}_i)^{-1}$. The first row contains w_{11i} and w_{12i}, the optimal weights for components one and two, respectively, for predicting the first criterion, while the second row contains w_{21i} and w_{22i}, the optimal weights for components one and two for predicting the second criterion. Given any value weights, the optimal weights for the composite is combinations of w_{11i} and w_{21i} for the first component and w_{12i} and w_{22} for the second component.

Focusing on the first criterion, algebraic manipulation reveals that

$$w_{11} = \frac{R_1 - R_1 R_2 \rho^2}{1 - R_1 R_2 \rho^2} \tag{4}$$

$$w_{12} = \frac{(1 - R_1) R_2 \rho}{1 - R_1 R_2 \rho^2} \tag{5}$$

Consider an extreme case where component 1 is completely reliable, $R_1 = 1$. This implies that $w_{11} = 1$ and $w_{12} = 0$, therefore all of the weight is put on component 1. In the opposite extreme, when component 1 is completely noise, $R_1 = 0$, then $w_{11} = 0$ and $w_{12} = \rho R_2$, and w_{12} equals the shrunken estimate of criterion 2, multiplied by the criteria regression coefficient. For any $0 < R_1 < 1$, w_{11} is decreasing in ρ and R_2, and w_{12} is increasing in ρ and R_2.

Next, suppose that component 2 is completely reliable, $R_2 = 1$. In this case $w_{11} = \frac{R_1 - R_1 \rho^2}{1 - R_1 \rho^2}$ and $w_{12} = \frac{(1-R_1)\rho}{1-R_1\rho^2}$. Taking the derivatives of the two equations with respect to ρ reveals that w_{11} is decreasing in ρ, and w_{12} is increasing in ρ. Taking the derivatives of the two equations with respect to R_1 reveals that w_{11} is increasing in R_1, and w_{12} is decreasing in R_1. In the other extreme, when component 2 is complete noise, then $w_{11} = R_1$ and $w_{12} = 0$, and the optimal weight for component 1 is the shrinkage estimate. For any $0 < R_2 < 1$, w_{11} is decreasing in ρ and increasing in R_1, and w_{12} is increasing in ρ and decreasing in R_1.

This example illustrates the intuitive result that as a component's reliability approaches 1, this component gets an increasing weight in the composite, and as the component's reliability approaches 0, this component receives less weight. It also illustrates that when criteria 1 is the target criterion, the weight on component 1 is small when the reliability of component 2 is large, and small when the correlation between the two criteria is large. Similarly, when the target criterion is criterion 1, the weight on component 2 is small when the reliability of component 1 is large and large when the

correlation between the two criteria are large.

4.2 Correlation with Target Criterion

Using this setup we can also examine the implications of reliability and target criterion correlation on the fit statistic defined in Equation 2, the correlation between the optimally weighted composite and the target criterion. Algebraic manipulation reveals that the fit statistic is increasing in the reliability of the components, and for a given reliability, it is increasing with the correlation among the criteria. The relative importance of the reliability of a component depends on the value weight of the component in the target criterion.

5 What is the Value of Additional Data?

A key question in the development of a composite is the value of different component measures for improving the prediction of a target criterion that was selected by stakeholders. We cannot assess the cost of data collection for individual districts, but we can assess the value of data for improving the predictions of the target criterion. Data can differ depending on the variables measured and the sampling plan for that measurement which will affect the size of the measurement error.

5.1 Data Reliability

The simple exercise above revealed that the reliability of the components determined the optimal weights and the correlation between the composite and the target criterion. Reliability depends largely on the data collection strategy and the type of measure being collected. For student level measures such as value added or student surveys, the reliability depends on the number of sections per teacher and the number of students per section. To improve the reliability of student level data, value added and survey information should be collected on all sections, and preferably from multiple years.

The reliability of teacher observation measures depends on a number of factors, such as the number of sections with lessons observed, the number of lessons observed, the number of ratings per lesson, and the number of raters per lesson. Because there is evidence that there is large rater-to-rater variability on teacher observation scores for a given lesson, higher reliability on teacher

observations can be achieved by rating multiple lessons with multiple ratings. In addition, adding multiple years of data will also improve teacher observation measure reliability.

5.2 Estimating the Value of Additional Data

To study the value of collecting more data for component measures, in the next version of this paper we will create alternative scenarios for data collection, varying the number of students, number of sections, and number of videos/raters. We will estimate $\boldsymbol{\mu}$ and the parameters of $\boldsymbol{A}$ and $\boldsymbol{E}_i$ via maximum likelihood with the data from the Measures of Effective Teaching project. We will determine the correlation between the composite and the target criterion for various possible composites including one based on optimal weights and one based on equal weights, and alternative target criteria including ones that use a single criterion, (e.g., student learning gains or teaching practices), and ones that combine individual criteria.[6] Finally, we will also examine the relative efficiency of weights estimated from regression methods that assume homoscedastic measurement error compared to truly optimal weights that account for heteroscedasticity in the measure across teachers.

The criteria that will be considered for this analysis are student achievement growth (estimated teacher VA) as a proxy for student learning, teacher observation protocol scores as a proxy for teaching techniques, and responses to student surveys as a proxy for student perceptions, satisfaction and ambition.

6 Discussion

This paper examines how to combine multiple measures of teaching into a single composite measure of teacher effectiveness. We discuss the need for experts and stakeholders to make value judgements on the target of the combined measure. We also derived the optimal composite for measuring the target criterion as the expected value of target given the observed component measures. The statistical weights are chosen to provide the best prediction of target criterion, and depend on the reliability of the components and the correlations among the criteria. Finally, we discussed the implications of different levels of the reliability of the components and the correlations among the

[6] The correlations will be conditional on our estimates of the underlying parameters determining $\boldsymbol{A}$ and $\boldsymbol{E}_i$.

criteria on the optimal statistical weights and the correlation between the composite and the target criterion.

It is straightforward for a district to implement the method described in this paper in two steps. The first step obtains the statistical weights for predicting individual criterion. These are obtained from a regression that predicts a teacher's performance on a given criterion in one year using all the measures of the teacher's performance from a different year. The regression coefficients from these regressions are the statistical weights, and the predictions from these regressions are the best linear prediction of each individual criterion. The second step uses value weights to combine the predictions of each individual criterion into an overall composite. The final composite is a weighted average of regression predictions of each individual criterion, where the weights are value weights determined by policy. Thus, rather than applying value weights to the individual measures (which is common practice), our method applies value weights to optimal predictions of each measure which account for the underlying measurement error in the data.

There are a number of important implications and limitations of our analysis. First, while a composite measure can provide a single summary, many experts argue against creating composites because they lose valuable information and can mask differences across measures. In addition, our analysis has assumed a low-stakes environment, and may not be optimal in high-stakes conditions. Finally, the calculations of optimal weights in this paper assumes that the components are unbiased estimates of valued criteria. If bias exists, then this bias needs to be built into the model, and will result in a different functional form for the optimal weights.

References

Behn, R. (2003). Why Measure Performance? Different Purposes Require Different Measures. Public Administration Review, 63(5), 586-606.

Carr, G. M., & Rickwood, C. J. (2008). Water Quality Index for Biodiversity Technical Development Document (Tech. Rep.). (Prepared for Biodiversity Indicators Partnership World Conservation Monitoring Center)

Dawes, R., & Corrigan, B. (1974). Linear Models in Decision Making. Psychological Bulletin, 81(2), 95.

Dimick, J., Staiger, D., Baser, O., & Birkmeyer, J. (2009). Composite Measures for Predicting Surgical Mortality in the Hospital. Health Affairs, 28(4), 1189-1198.

Editor, L. H. (2008). Rankings of Higher Education Institutions: A Critical Review. Quality in Higher Education, 14(3), 187-207.

Freudenberg, M. (2003). Composite Indicators of Country Performance: A Critical Assessment. OECD Science, Technology and Industry Working Papers.

Gulliksen, H. (1950). Theory of Mental Tests.

Holmstrom, B., & Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. Journal of Law Economics and Organization, 7, 24.

Johnes, G. (1992). Performance Indicators in Higher Education: A Survey of Recent Work. Oxford Review of Economic Policy, 8(2), 19–34.

Kane, T., & Staiger, D. (2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. The Journal of Economic Perspectives, 16(4), 91–114.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2010). The Worldwide Governance Indicators: Methodology and Analytical Issues. Policy Research Working Paper Series.

Klugman, J., Rodríguez, F., & Choi, H. (2011). The Hdi 2010: New Controversies, Old Critiques. Journal of Economic Inequality, 1–40.

Lehmann, E., & Casella, G. (1998). Theory of Point Estimation (Vol. 31). Springer Verlag.

Mehrens, W. (1990). Combining Evaluation Data from Multiple Sources. The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers, 322-334.

Morris, C. (1983). Parametric Empirical Bayes Inference: Theory and Applications. Journal of the

American Statistical Association, 47–55.

Murias, P., Miguel, J. de, & Rodríguez, D. (2008). A Composite Indicator for University Quality Assesment: The Case of Spanish Higher Education System. Social Indicators Research, 89(1), 129–146.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide (Tech. Rep.). (STI Statistics Working Paper)

Quality, N. C. T. (2011). State of the States: Trends and Early Lessons on Teacher Evaluation Effectiveness Policies (Tech. Rep.).

Reeves, D., Campbell, S. M., Adams, J., Shekelle, P. G., Kontopantelis, E., & Roland, M. O. (2007). Combining Multiple Indicators of Clinical Quality: An Evaluation of Differeny Analytic Approaches. Medical Care, 45(6), 489-496.

Rowena, J., Smith, P., & Goddard, M. (2004). Measuring Performance: An Examination of Composite Performance Indicators (Tech. Rep.). (Center for Health Economics Technival Paper Series 29)

Saisana, M., & Tarantola, S. (2002). State-of-the-art Report on Current Methodologies and Practices for Composite Indicator Development. EUR 20408 EN Report.

Schmidt, F., & Kaplan, L. (1971). Composite vs. Multiple Criteria: A Review and Resolution of the Controversy. Personnel Psychology, 24(3), 419-434.