A Composite Estimator of Effective Teaching

Kata Mihaly[1], Daniel F. McCaffrey[2], Douglas O. Staiger[3], and J. R. Lockwood[2]

[1]RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202-5050

[2]RAND Corporation, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213

[3]Dartmouth College, Hanover, NH 03755

January 8, 2013

Abstract

States and districts are collecting multiple measures of teaching to evaluate teacher effectiveness, but there is limited information about how indicators can be combined to improve inferences about a teacher's impact on student achievement and about teaching. We derive a statistical model and estimate the parameters of an optimal combined measure of teacher effectiveness using data from the Measures of Effective Teaching (MET) project. We contrast the optimal composites to composites created using equal weighting of indicators and to weights based on existing state policies. Our explorations consider multiple scenarios for data collection to determine tradeoffs between collecting more data and combining multiple indicators to improve the accuracy of inferences. We find evidence that there is a common component of effective teaching shared by all indicators, but there are also substantial differences in the stable component across measurement modes and across some indicators within a mode. The implication from our model is that composites that place relatively equal weight on all indicators will tend to capture the component of effective teaching that is common across indicators. We also find that optimal weights strongly depend on the target criterion and the optimal predictor tends to put most of the weight on the indicator corresponding to the target criterion. Composites formed based on state policies are moderately to highly correlated with optimal predictor of teacher contributions to achievement on the state test. Due to the relatively high reliability of the indicators in the MET project dataset, there are small differences in composites created under different data collection scenarios.

Note: This paper details the technical methods and analyses for the MET project's study of composite measures of teaching. A non-technical summary is in the MET project brief, "Ensuring Fair and Reliable Measures of Effective Teaching," available at www.metproject.org.

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, states and large school districts are rapidly developing and adopting new teacher evaluation plans. According to a report on teacher evaluation and effectiveness policies produced by the National Council of Teacher Quality, 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 indicators of effective 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 indicators of effective 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, compensation, and retention decisions.

Although states are developing complex rules for combining multiple indicators into a single composite score, there is limited empirical research to guide policymakers on how best to combine indicators to achieve specific goals or on the properties of such composite scores. This paper explores ways to combine multiple indicators of effective teaching to obtain optimal predictions of target criteria representing valued outcomes. It contrasts this approach to other methods of combining measures such as equal weighting and weighting according to policy priorities. We explore the properties of different ways for combining measures under different schemes for data collection using data from the Measures of Effective Teaching project.

2 Background and Conceptual Framework

Composite scores 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 (Holmstrom 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 scores 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 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 the 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. Similarly, the balanced scorecard approach (Kaplan and Norton (1992) was advocated for businesses 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 scores of teacher quality. Although some

states have developed complex rules for combining multiple indicators into a single composite score, there is tremendous desire for guidance on how to combine indicators to achieve specific goals. Two central questions arise in creating composites: 1) What dimensions of teaching do stakeholders and experts value and how are those characteristics to be measured? 2) What are the optimal statistical weights for predicting teacher performance on these dimensions and how are they determined from the data? In the remainder of this section, we address each set of questions in turn.

2.1 Value Decisions

The single composite measure of teacher performance will need to serve many purposes. States and districts are considering using the measure to support decisions for tenure, retention, and compensation. It will also direct professional development allocations and serve as a signal of a teacher's performance to guide the principal on how to improve the performance of the individual teacher and the school as a whole. The ideal composite indicator will lead to retaining the best teachers, to allocating professional development to teachers efficiently to obtain the maximum improvement in performance, and to supporting individual evaluations made by their supervisors.

This ideal will be challenging to achieve. Society values education for multiple reasons, such as helping citizens make positive contributions to the economy or achieving personal fulfillment. A teacher's contribution to each goal may not be equal. Experts and stakeholders will need to determine what it would mean to retain the "best" teachers. As noted in the OECD handbook on teacher evaluation (Nardo et al. (2008)), the first step to creating a composite score involves defining the underlying concept to be measured. It may be that no single concept is agreed to by all stakeholders and experts. In this case, the single composite measure will need to support decisions with good outcomes in multiple dimensions.

2.2 Optimal Predictions and Statistical Weights

When combining observed indicator measures into a composite one of three approaches may be used: the conjunctive, disjunctive, and compensatory models (Gulliksen (1950), Mehrens (1990)). The conjunctive model creates multiple cutoffs on each indicator 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 indicator meeting a threshold. For instance, in some 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 indicator to compensate for low values in other ones. The canonical compensatory model is the linear additive model which sets the composite equal to a weighted sum of the indicator measures. In this paper, we consider composites which are weighted sums of the indicators and the determination of those weights.

Mehrens (1990) distinguishes between a "clinical" and "statistical" approach to combining indicators to create a composite. With the clinical approach, judgment is used to choose weights for combining data. With the statistical approach, fixed weights are used to combine the data and statistical analysis 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)).

One approach to obtaining statistical weights is to assume that there exists a scalar target quantity, a "target criterion," which if known would be used to support all decisions and other functions of the composite measure. Decisions made using the target criterion would be the best available and the goal of the composite is to weight the indicators so that the resulting composite aligns very closely with the target criterion.

Provided such a target exists, then there exists weights which make the composite closest to the target (details are presented below). The resulting composite is the *optimal predictor* of the target because it is closest to it and the weights are the *optimal* or *statistical weights*. If the goal is to create a composite so that decisions made using the composite align most closely with those based on the target, then states will want to use the statistical weights and the optimal predictor. With the appropriate data, optimal weights can even be calculated from observed data (see the appendix for an example of how this might be achieved).

Choosing weights is unlikely to be as straightforward as the summary suggests. Teaching is likely to be multidimensional and different stakeholders might have different ideas about which dimensions

to target. Thus, there may be multiple target criterion. In this case we would need a composite that is "close" to many targets and we would need to develop a criteria for selecting the composite that is closest to multiple targets as well as considering composites that align with just one target. In addition, experts and stakeholder might not be able to fully explicate their targets, which would require a method for picking weights that perform well against a partially specified target or against a set of targets which are not fully defined. It will not be possible to find the composite that is closest to an undefined target. Consequently states might use the ideas of statistical weights to guide experts in the selection of weights that expert opinion suggests would perform well for predicting a partially specified target or set of targets rather than truly selecting statistical weights.

2.3 Objectives

In this paper we explore optimal weights for a restricted set of targets under different scenarios for data collection. Our goal is to understand how indicators like those that are being considered by states might be combined to improve inferences about a teacher's impact on student achievement (as measured by the state achievement test or by other tests) and about teaching (as measured by observations and surveys). We want to know how different indicators would contribute to these composites and how combining indicators affects the accuracy of our inferences about teachers and teaching. Although states or districts might not have single target, we use the context a single target criterion to understand the implications of different weighting scenarios. We also want to use these results to explore how indicators could be combined for predicting other target measures that might be considered by states and that might have unobserved but predictable relationships with value-added or teacher practices. As part of our exploration, we contrast the optimal composites for our limited targets to composites created using equal weighting of indicators or "policy" weights which place half of the weight on value-added and the remainder of the weight on the other indicators or half the weight on observations and the remainder on the other indicators. Our explorations consider multiple scenarios for data collection to determine tradeoffs between collecting more data and combining multiple indicators to improve the accuracy of inferences.

In the next section we describe the data used in our empirical analyses. The data description is followed by the development of our statistical models which formalizes some of the notions described

above. We then present results and end with a discussion of our findings and their implications for states and districts considering how to combine multiple measures of teaching.

3 Data

We use data from the Measures of Effective Teaching (MET) project to assess methods to combine data into composite measures of teaching. The MET project is a multi-year study of teaching performance measures supported by The Bill & Melinda Gates Foundation in six large school systems.[1] The study evaluated the performance of teachers using multiple measures, which we call indicators, including VA on the state accountability tests, VA on project-administered alternative assessments, student survey responses, and assessments of video recordings of classes. The study assessed teacher performance for two school years and randomly assigned class rosters to teachers for the second year to allow for testing VA on randomly assigned classes against VA on classes assigned using the schools' standard practices.

Our data collection design provided measures at the level of individual classrooms of students, so that most teachers have multiple measures of the same construct taken across multiple classrooms. The data are hierarchical with students (or video recordings) nested within classrooms, nested within teachers.

The MET project included grade four to eight mathematics and English language arts (ELA) teachers and high school teachers of English 1, Algebra 1, or biology. We restricted the sample to self-contained elementary school teachers and middle school teachers. For elementary school teachers, multiple classrooms were defined by two self-contained classrooms in two consecutive years. In addition, elementary school teachers were separately measured on the mathematics and ELA tests as if they were two separate classrooms. For example, both mathematics and ELA lessons were videotaped and scored by various classroom observation protocols, students responded to separate surveys about mathematics and ELA instruction, and VA measures were calculated separately for each subject. For middle school teachers, multiple classrooms were defined by two separate sections of the same subject from the same school year. For both elementary and middle school teachers, we

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

use the term "sections" to refer to the multiple classrooms on which measures were collected, even though in elementary schools they are self-contained classrooms from two school years.

3.1 Indicators

Value-added on the state accountability test

Each participating district provided the study with student achievement data on the state's mathematics and ELA accountability test for all students in grades 4 to 8. The districts also provided student scores on up to three prior years of testing. The ELA test covered reading, writing, and other aspects of ELA curriculum depending on the state. These data were used to construct VA measures at the level of sections. Because the tests from different states were on different scales, we standardize the scores by grade-level and subject within each state using the van der Waerden scores (Conover (1999)).[2] We refer to the van der Waerden scores as the standardized scores.

We estimated teacher VA using a two-step procedure. For the first step, by grade, subject and district, we modeled the standardized score as a linear function of the student's standardized score from the prior year on the same subject test, a set of student covariates, the average of the prior year test score on the same subject for the students in the section, and the percentage of students receiving free or reduced price lunches for students in the same section.[3] The student-level covariates used in the models included indicators for Black, Hispanic, age, gender, and English language learner, free and reduced price lunch, special education, and gifted statuses.[4]

We used least squares to estimate the model parameters and used the estimates to obtain fitted values and residuals for all students. In the second step, we averaged across all students linked to a teacher the residuals from the first stage to obtain the VA estimate. The second step included residuals from all grade levels for teachers who taught students in multiple grades (within the district and subject area). We refer to these scores as "state-test value-added" or SVA.

Value-added on the supplemental or alternative test

In addition to the state achievement tests in mathematics and English language arts, the MET

[2]To obtain the van der Waerden scores, scale scores were ranked within grade and subject by district and the percentile ranks were transformed to the quantiles of the standard normal distribution. The van der Waerden scores are similar to normal curve equivalents but scaled to a distribution with mean zero and standard deviation one.

[3]The value added model is described in detail in (Gates Foundation (2010)).

[4]The only exception was the Charlotte-Mecklenburg regressions which did not include information on free and reduced price lunch at the individual or classroom level because they were unavailable.

project administered supplemental mathematics and reading assessments. It administered three forms of the Balanced Assessment in Mathematics (BAM) in each of grades 4 to 8. Yearly, the BAM creates test forms for each grade-level with four to five open-ended tasks that require 50-60 minutes to complete. Because of the small number of tasks on each form the study administered the 2003, 2004, and 2005 forms each to roughly one third of students from each grade. For analysis, scores were standardized to van der Waerden scores by grade and form.

The MET project administered the Stanford 9 (SAT9) open-ended reading test to study students in grades 4 to 8. There were separate forms for each grade-level. Each form of the assessment consists of a narrative reading selection followed by nine questions. Students are required to not only answer the questions but also to explain their answers. The open ended nature of the questions distinguishes the SAT9 from traditional reading assessments. For analysis, scores were standardized to van der Waerden scores by grade and form.

We calculated VA estimates for both the BAM and SAT9 scores using analogous methods to those implemented with the state accountability tests. For the BAM we controlled for prior year state mathematics achievement scores and, for the SAT9, we controlled for prior year state English language arts test scores. We refer to these scores as "alternative-test value-added" or AVA.

Student perception survey

The MET project student perceptions survey is based on a survey developed over a decade by the Tripod Project for School Improvement. The Tripod questions are gathered under seven headings, or constructs, called the Seven C's: Care, Control, Clarify, Challenge, Captivate, Confer and Consolidate. Each of the C's is measured using multiple survey items and the items are combined according to the methods developed Ferguson (Gates Foundation (2010)) and confirmed through factor analyses conducted by the MET project. Lists of items used to measure each of the Seven C's can be found in (Gates Foundation (2010)). The Cronbach's alphas for the Seven C's are high, in the range of 0.80 and above. We combined the Seven C scales into a single composite equal to the average of the seven scales (referred to as student survey composite or SSC).[5] The scores from each student in the section taught by the teacher were averaged to obtain the section score. We centered the section average SSC scores to have mean zero in each district and removed peer

[5]The use of the combined seven scales is justified because preliminary analysis indicated that the individual scales are highly correlated and school districts are expected to use a combined measure.

characteristics by modeling the centered score as a linear function of the average of the prior year test score on the same subject for the students in the section, and the percentage of students on free and reduced price lunch for students in the same section.[6] We used least squares to estimate the model parameters and used the residuals as our measure of SSC.

The student perception survey also included three items on students effort in learning.[7] Exploratory factor analysis suggested these items constituted a measure of a uni-dimensional construct so we averaged them together to create a single measure of student effort (referred to as EFFORT or EFF). Similarly, the student perception surveys included questions on on how happy the student was in the classroom.[8] This student outcome is referred to as HAPPY or HIC for "Happy in Class". The EFFORT and HAPPY variables were averaged within sections, and peer characteristics were removed similar to procedures described for SSC.

Classroom observations

The MET project also evaluated teaching using four different observational protocols applied to video recordings of classes. Each teacher was asked to record four classroom periods for each subject they taught as part of the MET project. Secondary teachers were asked to record two lessons from each of two sections of classes. Elementary teachers teaching self-contained classes were asked to provide video recordings of four class periods of English language arts instruction (including reading if applicable) and video recordings of four class periods of mathematics instruction.

Raters scored each video recording using three different observation protocols. All video recordings were scored using the Classroom Assessment Scoring System (CLASS) and the Framework for Teaching (FFT) which are both subject matter neutral scoring systems. The video recordings were also scored using subject specific protocols: observers assessed English language arts classes using the Protocol for Language Arts Teaching Observations (PLATO) and mathematics classes were scored using the Mathematics Quality of Instruction (MQI).[9] Details on the content of the protocols

[6] We adjusted raw scores for peer characteristics because the value-added measures were adjusted for peers and we did not want peer effect to reduce the correlation between our measures.

[7] The three items are: (1) "When doing schoolwork for this class, I try to learn as much as I can and I don't worry about how long it takes;" (2) " I have pushed myself hard to understand my lessons in this class;" and "I have done my best quality work in this class."

[8] Both elementary and middle school surveys included the item "This class is a happy place for me to be" and elementary school surveys also included the reverse code of the item "Being in this class makes me feel sad or angry."

[9] The MET project modified PLATO and MQI, restricting some of the elements evaluated and developing new rater training materials. For details, see Kane et al (2012).

and the extent of external rater training are described in Kane et al. (2012).

Here we describe in detail the two protocols that are the focus of our analysis. FFT was developed from the Danielson teacher evaluation framework and uses a constructionist view of student learning with emphasis on intellectual engagement. The full framework includes domains outside of classroom observations but the MET project assessed only the two domains assessed via observation: classroom management and instruction. FFT has four component measures used to assess each of these two domains and we use the average of the eight component scores. The section-level score was calculated by taking the average of the ratings from multiple raters and multiple video recordings of the section.[10] These scores were centered to have mean zero at the district and grade level. We removed peer characteristics by modeling the centered scores as a linear function of the average of the prior year test score on the same subject for the students in the section, and the percentage of students on free and reduced price lunch for students in the same section. We used least squares to estimate the model parameters and used the estimates to obtain residuals.

PLATO focuses on instructional practices specific to language arts instruction including scaffolding through teacher modeling, teaching or English language arts strategies, and guided practice. The version of PLATO used by the MET project assesses teaching on six elements of instruction and we use the scores from these six elements. PLATO scores were averaged across ratings and videos, centered to have means zero with district and grade-level, and peer characteristics were removed similar to the procedure described above for FFT.[11]

4 Methods

In this section we present our statistical model for the observed the MET project indicator data and then describe our methods for using this model to study the prediction of various unobserved quantities from teachers.

[10] The combined measure is justified for observations since preliminary analysis suggested one factor and we expect districts to combine domains into a single measure.

[11] A similar procedure was used for CLASS and MQI. Details available on request.

4.1 Model for Indicator Data

We let y_{ijk} equal the observed values for the $k = 1, \ldots, K = 8$ *indicators* collected on each of $j = 1, 2$ sections from each of $i = 1, \ldots, N$ teachers. The models were estimated separately for four groups of teachers: ELA elementary (N=837), ELA middle (N=498), mathematics elementary(N=799) and mathematics middle (N=458). The indicators are value-added of the state test (SVA), value-added on the alternative test (AVA), ratings on the Framework for Teaching (FFT), CLASS (CLASS), and PLATO or MQI (PLATO/MQI) depending on the subject of instruction, composite scores from the Seven C indices from the supplemental student survey (SSC), and adjusted mean values for the effort (EFF) and happy in class (HIC) indices from the surveys.

Associated with each indicator is a *stable component* of the measure, ϕ_{ik}. As discussed in the MET project reports (Gates Foundation (2010), Kane et al. (2012)), the stable component is a teacher's average performance over a longer period of time and multiple classes. It is similar to a universe or true score in a measurement model or generalizability theory (Allen and Yen (2001), Brennen (2001)). By definition, each indicator is an unbiased estimator of the associated stable component so that

$$y_{ijk} = \phi_{ik} + \epsilon_{ijk}, \tag{1}$$

where the ϵ_{ijk} are the measurement or sampling error in the indicators, with $E(\epsilon_{ijk}) = 0$, and ϵ_{ijk} independent across teachers and independent of ϕ_{ik}. The ϵ_{ijk} may be correlated across indicators from the same section due to measures on the same students or class. Each ϵ_{ijk} consists of two sources of error:

$$\epsilon_{ijk} = \zeta_{ijk} + \alpha_{ijk}, \tag{2}$$

where the ζ_{ijk} is variation due to aggregating measures across students or videos within a class or section, and the α_{ijk} is variation due to the sections or classes taught by the teacher. We assume these two sources of error are all mean zero and independent within teacher and across teachers but they can be correlated across indicators.

We let $\boldsymbol{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijK})'$ be a vector of indicators for section j taught by teacher i, and $\boldsymbol{\phi}_i$, $\boldsymbol{\epsilon}_{ij}$, $\boldsymbol{\alpha}_{ij}$, and $\boldsymbol{\zeta}_{ij}$ be the corresponding vector of stable components, and measurement errors and their component sources. We let $\boldsymbol{\mu}$ equal the mean of $\boldsymbol{\phi}_i$ and $\boldsymbol{A}$, $\boldsymbol{E}_{ij}$, $\boldsymbol{B}$ and $\boldsymbol{S}_{ij}$ equal the variance-

covariance matrices for $\boldsymbol{\phi}_i$, $\boldsymbol{\epsilon}_i$, $\boldsymbol{\alpha}_{ij}$, and $\boldsymbol{\zeta}_{ij}$, respectively. By assumption $\boldsymbol{E}_{ij} = \boldsymbol{B} + \boldsymbol{S}_{ij}$. Both the stable components and the measurement errors for a teacher can be correlated across indicators so each variance-covariance matrix may have nonzero off-diagonal elements. The variances and covariances among the ζ_{ijk} will depend on the number of students or videos assessed in the section for each indicator. Hence, $\boldsymbol{S}_{ij}$ will vary across sections and teachers. We use the MET project data to estimate the indicator means and the variance covariance matrices. We assume that $\boldsymbol{\phi}_i$, $\boldsymbol{\alpha}_{ij}$, and $\boldsymbol{\zeta}_{ij}$ are multivariate normal vectors.

4.1.1 Teacher-level Data

The MET project data are section-level which is necessary for estimating the parameters of the $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices. However, teacher evaluations will rely on teacher-level data that combine the data from multiple sections into a single measure vector of indicators for each teacher. The most precise teacher-level indicators will equal weighted sums of the section-level indicator where the weights equal the reciprocals of the diagonal elements of $\boldsymbol{E}_{ij}$, the variances of the indicators among sections from the same teacher. If the error variance is constant for a teacher, then the weights are constant and the teacher-level indicators $\boldsymbol{Y}_i = n_i^{-1} \sum_j \boldsymbol{Y}_{ij}$ and the variance-covariance matrix for $\boldsymbol{Y}_i$ is $\boldsymbol{E}_{ij} = \boldsymbol{B}_i + \boldsymbol{S}_i$, where $\boldsymbol{B}_i = n_i^{-1}\boldsymbol{B}$ and $\boldsymbol{S}_i = n_i^{-2} \sum_j \boldsymbol{S}_{ij}$.[12] In the remainder of the paper we focus on the teacher-level indicators except when discussing the estimation of $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{S}_{ij}$.

4.2 Estimation

To estimate the parameters of our statistical model, we first estimated $\boldsymbol{S}_{ij}$ for each section. We then treated these values as known and estimated $\boldsymbol{\mu}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ by maximizing the likelihood for $\boldsymbol{y}_{ij}$. We use bootstrap resampling to estimate the standard error of the variance and covariance parameters of $\boldsymbol{A}$ and $\boldsymbol{B}$ and functions of these including correlations and reliabilities. Details on our estimation methods including the estimation of $\boldsymbol{S}_{ij}$ and the maximum likelihood estimators for the other parameters and our bootstrap resampling are provided in the appendix.

[12] If $\boldsymbol{S}_{ij}$ are not constant within a teacher because of different class sizes or numbers of ratings, then $\boldsymbol{Y}_i = \sum_j \boldsymbol{Y}_{ij}\boldsymbol{W}_{ij}$, where $\boldsymbol{W}_{ij}$ is a diagonal matrix of the weights for each indicator, and $\boldsymbol{B}_i = \sum_j \boldsymbol{W}_{ij}\boldsymbol{B}\boldsymbol{W}_{ij}$ and $\boldsymbol{S}_i = \sum_j \boldsymbol{W}_{ij}\boldsymbol{S}_{ij}\boldsymbol{W}_{ij}$.

4.3 Prediction

We consider composite estimators of the form of a weighted sum of observed indicators. We assume that the composite is meant to predict the unobserved target criterion. As discussed below, the accuracy of the composite for predicting the target criteria depends on $\boldsymbol{A}$ and $\boldsymbol{E}_i = \boldsymbol{B}_i + \boldsymbol{S}_i$. For teachers in the MET project $\boldsymbol{E}_i$ depends on the specific data collection protocol used in the MET project. While the data collection design used in the MET project was specifically chosen to support the research study, in education practice different designs might be used. Moreover, policymakers designing teacher evaluation systems are interested in knowing how different data collection plans for teachers in their states and districts might affect the properties of their teacher performance measures. Hence, rather than assess the properties of composite measures for the MET project teachers, we study the properties of alternative composite measures for a set of four data collection design scenarios.

For each scenario, we assume that states and districts will collect teacher-level indicators, $\boldsymbol{Y}_i$, that include a growth measure like value-added on the state test, a classroom observation and a third measure. We use SVA, FFT, and the survey composite, SSC, as representative measures for the data being used by states.[13] Because there are questions about the value of subject-specific observation protocols relative to a generic protocol like FFT, we also consider cases where the indicators are SVA, PLATO, and SSC. For data collection scenario, s, there are corresponding $\boldsymbol{B}_{[s]i}$, $\boldsymbol{S}_{[s]i}$, and $\boldsymbol{E}_{[s]i}$ matrices. The matrix $\boldsymbol{B}_{[s]i}$ equals the rows and columns of $\boldsymbol{B}$ for sections corresponding to SVA, FFT (or PLATO) and SSC but with the appropriate scaling of elements for the numbers of sections used to calculate each indicator. Similarly, the elements of $\boldsymbol{S}_{[s]i}$ equal the variance and covariance components for SVA, FFT (or PLATO) and SSC scaled by the number of students or videos assumed for scenario s. We present analysis in which each section has the same number of students and discuss cases with heterogeneous error variance in the appendix.

For each scenario, we use the estimates of $\boldsymbol{A}$, $\boldsymbol{B}$ and the $s_{kk'}$ components $\boldsymbol{S}_{ij}$ from our MET project sample to create $\boldsymbol{A}_{[s]}$, $\boldsymbol{E}_{[s]i}$, $\boldsymbol{B}_{[s]i}$, and $\boldsymbol{S}_{[s]i}$. We then use these values in the formulas for the "optimal" predictor weights described below to determine the weights given to SVA, FFT (or PLATO), or SSC for predicting various quantities of potential interest to states. We also evaluate

[13] Informal reports from states suggested that several were using FFT or a modification of that protocol.

each predictor using the fit statistics described below to determine how well we can predict a specified quantity of interest and the robustness of a prediction for predicting other quantities that may be of interest.[14]

4.4 Data Collection Scenarios

To explore the quality of prediction of the target criterion from various data collection strategies, we devised a number of scenarios that approximate real world data collection options faced by schools. We considered twelve scenarios. These scenarios vary the amount of data collected for calculating SVA, FFT, and SSC. The scenarios are structured as four scenarios per number of years of data collection, where we assumed either one, two, or three years of data used in the predictions. The scenarios for a given number of years of data differ for elementary and middle school teachers, as described below. In the analysis we focus on estimates using the highest level data collection plan for one year of data collection, and contrast results from this plan to adding years of data collection or lowering the quality of data for a single year.

Middle School Scenarios

We assume that in middle schools the teacher has 20 students in each of four classroom sections. Twenty is the median number of students per section used to estimate student achievement VA on the classroom rosters collected by the MET project, and four sections represent about the average number of sections we have found in data.[15]. All of the students in the teacher's classroom are assumed to contribute to the teachers' VA on the state test for each of the scenarios. The quality of the value added data is therefore only contingent on the number of years of data collected (one, two, or three).

Next we consider the quality of data on student surveys and teacher observations. In our primary data collection plan with our higher levels of data collection for observations and surveys, (VID:high, SSC:high), the student survey data are assumed to be collected from all four sections taught by the teacher and the teacher is assumed to be observed four times, two times by one rater and two times by a different rater. In the remaining scenarios: the student survey data are assumed to be collected

[14]To simplify the notation we will drop the "[s]" subscript in the remainder of the paper.

[15]In addition, the average number of sections from an alternative study with very high quality data was also four (www.utqstudy.org)

from only two of the four sections taught by the teacher (VID:high, SSC:low); the teacher is assumed to be observed one time by one rater (VID:low, SSC:high); or only one section is surveyed *and* the teacher is observed only one time (VID:low, SSC:low).

Elementary School Scenarios

For elementary school teachers we assume the teacher teaches a self-contained class of 20 students. Again, 20 is the median number of students used to estimate VA on the classroom rosters collected by the MET project. All of the students in the classroom are assumed to contribute to the teachers' VA on the state-test for all scenarios. Therefore the value added data quality for elementary schools also depends on only the number of years of data collected.

In our primary data collection plan with our higher levels of data collection for observations and surveys, (VID:high, SSC:high), the student survey data are assumed to be collected for both subjects from all students taught by the teacher and the teacher is assumed to be observed four times, two times by one rater and two times by a different rater. In the remaining scenarios: the student survey data are assumed to be collected for each subject (mathematics or ELA) from only half of the students taught by the teacher (VID:high, SSC:low); the teacher is assumed to be observed one time by one rater (VID:low, SSC:high); or only half of the students complete the survey for each subject *and* the teacher is observed only one time (VID:low, SSC:low),

4.5 Optimal Weighting

We assume that states and districts are interested in a target criterion, a univariate quantity of interest that if known would be the preferred value for making the decisions about teachers that the composite performance measure will be used to support. We denote the target criterion by η_i. One possible approach for creating a composite would be to choose the combination of scores that is most correlated with the target criterion or closest to it in terms of expected squared difference (mean squared error, MSE) between the unobserved target criterion and a composite measure. We call this the "optimal predictor" because it minimizes MSE and maximizes the correlation between the composite and the quantity of interest.

In general, the optimal predictor must equal the expected value of the target criterion given the observed indictors, under the statistical model for the data (Lehmann and Casella (1998)).

When the data can be modeled using a multivariate normal or other elliptical distribution, then the optimal predictor for the target criterion will be a weighted sum of the indicators. This composite will have the highest correlation with the target criterion among all linearly additive (weighted sum) predictors of it.

Assuming the indicators are multivariate Gaussian or otherwise follow an elliptical distribution, then

$$E[\,\eta_i \mid \boldsymbol{Y}_i\,] = a + \boldsymbol{c}'(\boldsymbol{A} + \boldsymbol{E}_i)^{-1}\boldsymbol{Y}_i \tag{3}$$

where a accounts for the mean of the target and the indicators, and $\boldsymbol{c}$ is vector of length K with elements $cov(\eta_i, \phi_{ik})$.[16][17][18]The optimal weights are the population regression coefficients of a regression of η_i on the indicators. If the variance-covariance matrix of the measurement errors varies across teachers the weights will also vary. We use the $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{S}_i$ matrices derived from the MET project sample to calculated the statistical weights for the optimal predictors.

4.6 Evaluation Criteria

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 assess potential composite measures by their correlation with the target criterion and their stability across years. As noted by Kane and Staiger (2002) the correlation between the target and the composite measure determines the accuracy of the measure. The stronger the correlation the more accurate the measure. A strong correlation means that decisions made using the composite will be similar to those made using the targets. When the correlation is strong there will be fewer teachers misclassified as being effective or ineffective as measured by the target and the ranking of teachers will be very similar to what it would be on the target. When the correlation is low, classifications made on the composite will misclassify many teachers and rankings will not align with those from the target.

[16]Our assumptions imply that the estimator is the best linear predictor of η_i. If the multivariate Gaussian or elliptical distribution assumption is violated then the estimator will remain the best linear predictor, but not the best overall predictor. The estimate of the correlation matrices are robust to violations of the distributional assumption.

[17]Equation 3 can yield negative weights which may be undesirable. We can find optimal weights which are constrained to be greater than zero by finding the weight vector $\boldsymbol{w}$ which minimizes $E[(\,\eta_i - \boldsymbol{w}'\boldsymbol{Y}_i\,)^2]$ subject to the constraint that the elements of $\boldsymbol{w}$ are all greater or equal to zero. If the solution to Equation 3 has elements that are greater than or equal zero it minimizes the alternative equation.

[18]We standardize our optimal weights to sum to one across indicators.

Formulas for the correlation are provided in section B of the appendix.

Assuming that teacher evaluations will focus on stable effects of the teacher, instability in composite measures across years will mask true teacher effectiveness and lead to confusion for teachers and school administrators. Consequently, all else being equal, composites that are more stable across years will be preferable to ones with larger year-to-year variability, and we consider stability of the composite as part of our evaluation. Formulas for stability are also provided in section B of the appendix.

5 Results

The primary goal for any composite of teaching indicators is to support decision making on teachers. To achieve this goal, we want to specify weights so that the weighted sum of the indicators is closest to the underlying measures of teaching that would allow for error free decisions on teaching. We also want composites that will be stable across time in cases where teacher performance does not change. This will avoid sending teachers mixed signals when their performance does not vary year to year.

Two factors will determine how well any single composite can achieve these goals:

1. the measurement error in the indicators
2. the correlation structure among the stable components associated with the indicator measures.

In this section, we first present estimates of the measurement errors and the correlation among the stable components in the MET project indicators. We then use these estimates to construct optimal statistical weights for predicting the stable component associated with each of the eight indicators measured by the MET project. Finally, we evaluate the performance of these and other weighting schemes in terms of the stability of the resulting composite indicators and their correlation with the stable component of selected MET project indicators.

5.1 Estimates of Measurement Error

There are two sources of error variance in indicators: the section-to-section variability and the variability resulting from aggregating measures from individual students or lessons. Table 1 reports

estimates of the section-to-section variability in each of the eight MET project indicators. For easier interpretation, we report the standard deviation of the section-level error relative to (i.e. divided by) the standard deviation of the stable component. For all of the student-based indicators (value-added or survey indicators), the section-to-section variability is large relative to variability in the stable component, with the standard deviation in the section components ranging from 50 to 130 percent of the standard deviation in the stable components. Generally the ratio of section-to-section variability to the stable component variability is larger in elementary schools than in middle schools. The variability across sections in these measures exists even though our value-added and survey measures control for the section level average prior achievement and socio-economic status (percent FRL eligible students). These estimates suggest that teacher performance on the student-based indicators varies considerably across sections, particularly in elementary grades, perhaps due to idiosyncratic factors such as classroom chemistry or peer effects.

Interestingly, the section to section variability in the classroom observations is very small. Teaching practice and classroom interactions in the video recorded lessons demonstrates little to no variability among the sections for both the elementary and middle school teachers in our sample. Apparently, classroom observations for a teacher are fairly insensitive to the particulars of a given classroom.

The reliability of our indicators presented in Table 2 depends on both sources of error variance. We focus on the data scenario with one year of data collection and high quality data on observations and surveys. While the relatively low reliability of value added scores in elementary grades has been well documented, Table 2 suggests that classroom observation and student survey indicators have similar levels of reliability. Middle school indicators are more reliable than elementary school indicators. For value-added and survey indicators this is because they include more sections and more students. However, middle school observations are also more reliable than elementary school observations even though they include the same number of observations made by the same number of raters. This is due to the greater variability among the stable components of middle school teachers, especially on FFT. Mathematics indicators are more reliable than ELA indicators for value-added, because the stable component variability is larger from mathematics tests than ELA tests.

Errors in inference made on the composite can be reduced by putting more weight on measures

Table 1: **Standard Deviation of Section Level Effects Divided By Standard Deviation in Stable Components for MET Project Indicators**

	Elementary		**Middle**	
	Math	**ELA**	**Math**	**ELA**
SVA	0.78	1.02	0.50	1.21
	(0.08)	(0.20)	(0.07)	(0.26)
AVA	1.05	1.20	1.06	0.73
	(0.16)	(0.25)	(0.20)	(0.09)
CLASS	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
FFT	0.26	0.21	0.16	0.31
	(0.08)	(0.06)	(0.10)	(0.10)
MQI/PLATO	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
SSC	0.75	0.95	0.55	0.58
	(0.14)	(0.14)	(0.05)	(0.06)
EFF	1.30	1.19	0.87	0.60
	(0.31)	(0.29)	(0.12)	(0.10)
HIC	0.83	1.13	0.63	0.60
	(0.18)	(0.21)	(0.07)	(0.06)

Note: Standard errors of the estimated standard deviations are in parentheses. Section level variance on observation scores for CLASS, MQI and PLATO were set to zero on basis of initial data explorations.

Table 2: **Reliabilities of Predictors from 1 Year, FFT - High, SSC - High Data Scenario**

	Elementary		**Middle**	
	Math	ELA	Math	ELA
SVA	0.50	0.32	0.85	0.46
	(0.04)	(0.06)	(0.02)	(0.08)
AVA	0.33	0.29	0.69	0.80
	(0.05)	(0.06)	(0.06)	(0.03)
FFT	0.45	0.40	0.67	0.68
	(0.03)	(0.03)	(0.03)	(0.03)
CLASS	0.36	0.30	0.56	0.58
	(0.03)	(0.02)	(0.03)	(0.02)
MQI/PLATO	0.34	0.34	0.24	0.58
	(0.03)	(0.03)	(0.06)	(0.03)
SSC	0.51	0.41	0.90	0.89
	(0.07)	(0.06)	(0.01)	(0.02)
EFF	0.15	0.20	0.68	0.77
	(0.05)	(0.06)	(0.04)	(0.04)
HIC	0.41	0.30	0.85	0.86
	(0.07)	(0.06)	(0.02)	(0.02)

Note: Standard errors of reliability estimates are in parentheses

that are more reliable. For instance, it can be shown algebraically that the optimal weight for an indicator for predicting a target criterion will increase proportionately with square root of the indicator's reliability. The ranges in the square roots of the reliabilities for the various indicators for each subject and grade level are modest and, therefore, reliability will have a modest impact on the determination of the optimal weights. It can also be shown algebraically that optimal weights are proportional to the correlation between the stable components and the target indicator. Because reliability is similar across our indicators, the weights will depend primarily on differences in the correlation between stable components and the target criteria. We now turn to the correlations among the stable components and their implications.

5.2 Correlation Among Stable Components

Table 3 presents the estimates of the correlations among the stable components associated with each of the MET project indicator measures. To simplify the table, we report the correlation between each of the eight MET project indicators with three selected indicators representing each major

measurement "mode": SVA representing the test-based indicators, FFT representing the teacher observation indicators, and SSC representing the student survey indicators.

In general, the stable component of all the 8 MET project indicators are correlated positively, indicating that they measure some common dimension for teachers. There is higher correlation within measurement mode (these correlations are shown in bold in the table), namely teacher observation protocol scores are highly correlated with one another, and student survey responses to the SSC are highly correlated with responses to the EFF and HIC survey items. The common mode effect appears to be the strongest for teacher observations and student perception survey responses, and much weaker for the test-based indicators (where the correlation is between 0.39 and 0.54). This low correlation between the stable components for tests indicates that there is a sizeable unique component associated with each of the test-based measures, a component that is not captured by the impact of the teacher on scores on the other test. There is something common to tests that is not shared by the other measures, just like there are common components for observation and survey measures, but the unique components are relatively larger and the common component is relatively smaller for tests than for observation and surveys.

Overall, Table 3 suggests that there is a common component of effective teaching shared by all indicators, but there are also substantial differences in the stable component across modes and (for value added) across indicators within a mode. This has important implications for composites. Composites that place relatively equal weight on all indicators will tend to capture the component of effective teaching that is common across indicators and mode, and therefore be a good predictor of all dimensions of teaching. In contrast, composites that place more weight on a particular indicator will tend to capture more of the components of effective teaching that are unique to that indicator and mode, and therefore be a better predictor of teacher performance on that particular indicator but a worse predictor on other indicators. In the remainder of this section, we use our estimates to more formally evaluate the properties of composites using alternative weighting schemes.

Table 3: **Estimated Correlations Among Stable Components**

	Elementary						**Middle**					
	Math			**ELA**			**Math**			**ELA**		
	SVA	FFT	SSC	SVA	FFT	SSC	SVA	FFT	SSC	SVA	FFT	SSC
SVA	1.00	0.27	0.33	1.00	0.28	0.50	1.00	0.41	0.44	1.00	0.17	0.29
	—	(0.07)	(0.07)	—	(0.10)	(0.12)	—	(0.08)	(0.06)	—	(0.13)	(0.12)
AVA	**0.43**	0.10	0.19	**0.54**	0.35	0.33	**0.39**	0.42	0.32	**0.41**	0.30	0.16
	(0.10)	(0.09)	(0.11)	(0.12)	(0.09)	(0.12)	(0.09)	(0.10)	(0.09)	(0.17)	(0.10)	(0.08)
CLASS	0.28	**0.98**	0.56	0.28	**0.99**	0.43	0.35	**0.89**	0.57	0.32	**0.89**	0.51
	(0.06)	(0.01)	(0.08)	(0.10)	(0.01)	(0.09)	(0.09)	(0.04)	(0.07)	(0.13)	(0.03)	(0.08)
FFT	0.27	1.00	0.43	0.28	1.00	0.41	0.41	1.00	0.50	0.17	1.00	0.45
	(0.07)	—	(0.08)	(0.10)	—	(0.09)	(0.08)	—	(0.07)	(0.13)	—	(0.08)
MQI/PLATO	0.19	**0.53**	0.42	0.34	**0.86**	0.21	0.42	**0.80**	0.45	0.35	**0.92**	0.34
	(0.09)	(0.06)	(0.08)	(0.11)	(0.05)	(0.10)	(0.12)	(0.09)	(0.12)	(0.14)	(0.04)	(0.08)
SSC	0.33	0.43	1.00	0.50	0.41	1.00	0.44	0.50	1.00	0.29	0.45	1.00
	(0.07)	(0.08)	—	(0.12)	(0.09)	—	(0.06)	(0.07)	—	(0.12)	(0.08)	—
EFF	0.54	0.46	**0.91**	0.45	0.32	**0.69**	0.57	0.51	**0.94**	0.41	0.40	**0.89**
	(0.12)	(0.12)	(0.07)	(0.17)	(0.12)	(0.10)	(0.09)	(0.09)	(0.03)	(0.14)	(0.10)	(0.02)
HIC	0.32	0.45	**0.92**	0.34	0.39	**0.88**	0.37	0.37	**0.94**	0.11	0.33	**0.96**
	(0.07)	(0.09)	(0.03)	(0.14)	(0.10)	(0.05)	(0.07)	(0.09)	(0.01)	(0.12)	(0.09)	(0.01)

5.3 Optimal Weights

Figure 1 displays the optimal statistical weights for predicting the stable component associated with each of the eight indicators measured by the MET project. These weights are chosen to maximize the correlation between the resulting composite and the stable component of each indicator. Given the current policy focus on improving student test scores, weights for predicting the stable component of teacher value added (SVA or AVA) are of particular interest. However, optimal weights to predict a teacher's stable component on teacher practices like those measured by FFT or surveys may be of interest for other purposes (such as professional development) and also illustrate how weights will differ depending on the target criteria.

We calculated separate weights by subject and grade level, and focus on the data quality scenario with one year of data collection, four teacher observations, and all of the teacher's students surveyed. This scenario was chosen because it closely resembles the data collection capacity of many districts around the country. It suggests the optimal weights under the maximum data collection. We later consider lower quality data collection scenarios. The weights were calculated using the standardized version of each indicator to ensure that the scale of the measures is comparable and does not drive the results.

One clear pattern that emerges is that the largest weight falls on the indicator that is within the same measurement mode as the target criterion, with the indicator from the same mode often getting more than 75 percent of the weight in the composite. Namely, the SVA indicator received the largest weight for predicting SVA and AVA, the FFT indicator receives the largest weight for predicting all of the video measures, and the SSC indicator receives the largest weight for predicting the student survey measures. This reflects the observation from Table 3 that mode effects are relatively large.

The weakest mode effect occurs when predicting the stable component of AVA, with a weight on SVA that is substantially smaller than the weight that SVA receives when predicting the stable component of SVA. For AVA we can only predict the common test mode component of the stable component. This is best predicted by SVA, which is from the same mode, but because mode component is small relative to the unique component in test scores SVA receives less weight (and FFT and SSC more weight) in predicting AVA. For FFT and the survey indicators, the unique components are relatively smaller than the mode components, as noted above, and consequently the

weight on FFT for predicting CLASS or the subject specific measure tends to be very similar to the weight on FFT for predicting stable FFT and the weight on SSC for predicting EFF and HIC are about as large as the weight for predicting stable SSC.

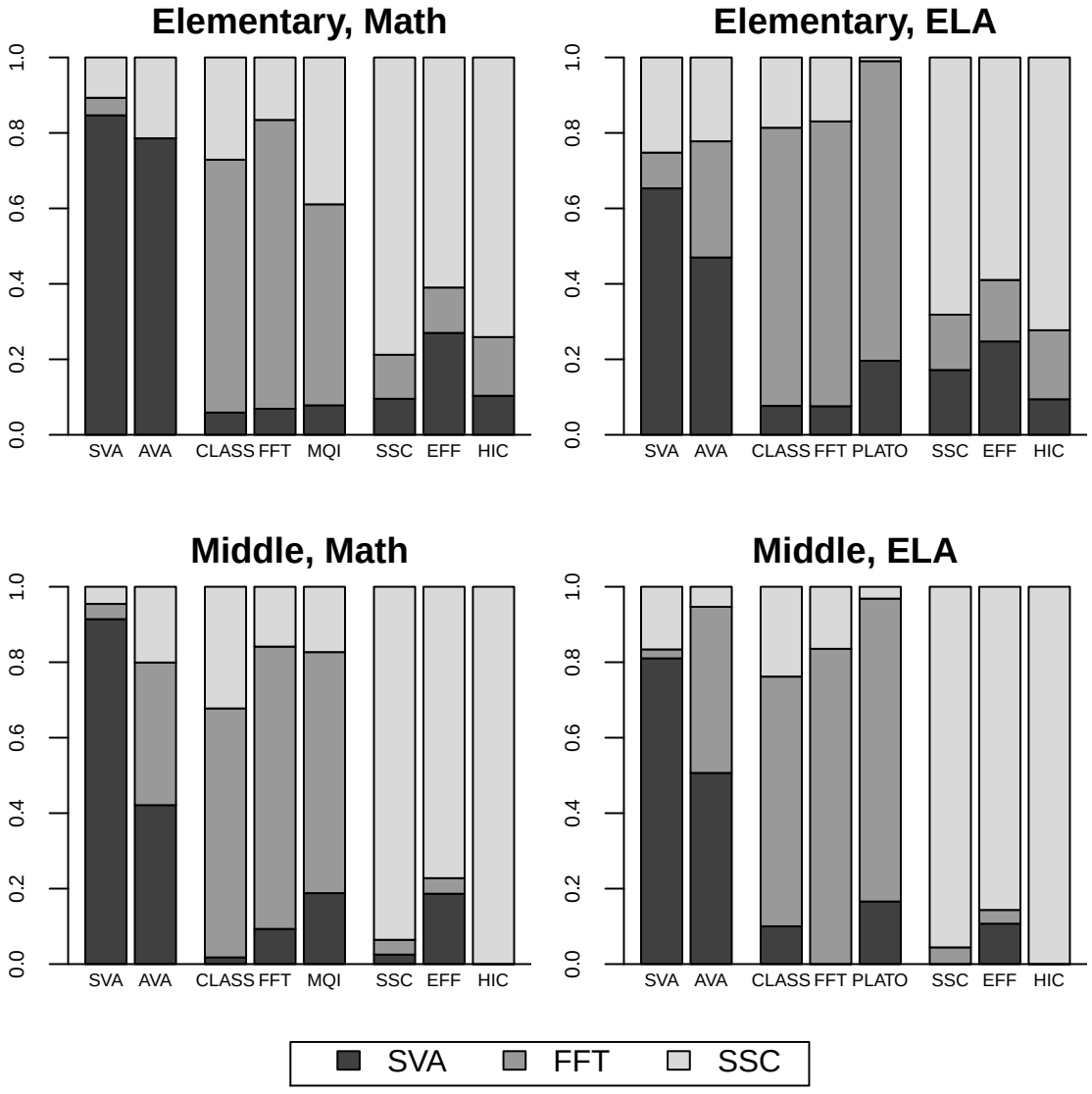
The weights are much more sensitive to the correlation of the target criteria than the reliability of the indicator. For instance, the reliability of FFT is much lower than the reliability of SVA or SSC for middle school mathematics teachers, but the vast majority of the weight for predicting CLASS or MQI remains on FFT, and the weight is nearly as disproportionate on FFT for middle school mathematics as it is for ELA or elementary school mathematics. Similarly, there is relatively small variation in the weights when different data scenarios are considered that vary the reliability of each indicator (results not shown). Measures with higher reliability receive somewhat larger weight, but weighting indicators in the same mode continues to dominate.

5.4 Evaluating Alternative Composites

As seen in Figure 1, the optimal statistical weights for a composite depend strongly on the target criteria. Policy makers, however, may disagree about what is the appropriate target. Rather than take a stand on what should be the appropriate target, in this section we evaluate a range of alternative composites that might be considered. We evaluate ten alternative combinations of SVA, FFT, and SSC: the best predictor of SVA, FFT, and SSC; each of the indicators alone (all of the weight on only one indicator); 50 percent weight on each of FFT and SSC; 50 percent weight on SVA with 25 percent weight each on FFT and SSC; 50 percent weight on FFT with 25 percent weight each on SVA and SSC; and equal weight on all three indicators. The last three scenarios correspond to current practice in many states and districts. We include scenarios without SVA to demonstrate predictions for teachers without student growth data. For some analyses we also consider the best predictor of AVA. Again, all indicators are standardized to have variance equal to one so that weighting does not depend on the scaling of measures.

The ordering of teachers will be sensitive to choice of the composite for any year, because the different combinations give very different weights to different indicators, yielding composite measures that are often only moderately correlated. Table 4 presents the correlation between the best predictor of SVA, a target that has been suggested in the past (Gates Foundation (2010)),

Figure 1: **Optimal Statistical Weights, by Subject and Grade Level.** Each bar presents the weights on SVA, FFT, and SSC for optimal predictors of the eight indicators collected by the MET project under a data collection scheme of 1 Year, FFT - High, SSC - High.



and the other composite measures. The best predict of SVA correlates very highly with SVA alone because as shown in Figure 1, the best predictor of SVA puts nearly all the weight (around 80 percent) on SVA. The best predictor of SVA correlates weakly with best predictors of FFT and SSC and FFT or SSC because those measures put most or all of their weight on either FFT or SSC, and very little weight on SVA. Similarly, the best predictor of SVA is weakly correlated with a 50-50 FFT-SSC measure. The correlation is larger with the 50-25-25 and equal weights because SVA tends to have larger weight in these composites.

Table 4: **Correlation of Best Predictor of SVA with Alternative Composite Measures**

	Elementary		Middle	
	Math	ELA	Math	ELA
Best predictor of FFT	0.37	0.46	0.50	0.26
Best predictor of SSC	0.45	0.72	0.48	0.40
SVA alone	0.99	0.93	1.00	0.98
FFT alone	0.25	0.29	0.36	0.20
SSC alone	0.33	0.53	0.45	0.40
FFT 50/SSC 50	0.36	0.53	0.48	0.37
SVA 50/FFT 25/SSC 25	0.90	0.97	0.90	0.91
SVA 25/FFT 50/SSC 25	0.64	0.73	0.70	0.62
Equal weights	0.76	0.87	0.79	0.76

As shown in Table 5, the stability of the composite across years is somewhat less sensitive to the choice of the composite. The stability is driven by the reliability of the indicators that receive the most weight. Combining measures improves the stability of the measure over a single indicator. On average across measures, subjects, and grade levels, the stability of the best predictor of SVA, FFT, or SSC is about nine percent greater than the stability of the single indicators. Equal weighting or policy weighting (50-25-25) generally yield the most stable measures except for middle schools where the best predictor of SSC is most stable because the surveys measures are very reliable with four sections of students surveyed. However, policy and equal weighting are nearly as reliable even for this intensive data collection plan.

5.4.1 Comparing correlation of various predictions with various targets

The first row of Table 6 shows that the stable component of each indicator can be predicted well by its optimal predictor, especially for middle school teachers. The correlation between the opti-

Table 5: **Year-to-Year Stability of Alternative Composite Measures**

	Elementary		Middle	
	Math	ELA	Math	ELA
Best predictor of SVA	0.52	0.42	0.86	0.51
Best predictor of FFT	0.49	0.44	0.74	0.72
Best predictor of SSC	0.54	0.48	0.90	0.89
SVA alone	0.50	0.32	0.85	0.46
FFT alone	0.45	0.40	0.67	0.68
SSC alone	0.51	0.41	0.90	0.89
FFT 50/SSC 50	0.54	0.48	0.84	0.84
SVA 50/FFT 25/SSC 25	0.57	0.46	0.88	0.66
SVA 25/FFT 50/SSC 25	0.55	0.49	0.83	0.75
Equal weights	0.57	0.50	0.88	0.76

mal predictor and the stable component ranges from .61 for SVA elementary ELA to .95 for SSC middle school mathematics. Thus, optimal composites are fairly strongly correlated with the stable components they are intended to capture.

The remaining rows of Table 6 report the correlation of alternative composites with the stable components of SVA, FFT and SSC. These are given as a percentage of the correlation with the optimal predictor in order to emphasize the loss from using alternative composites. Each indicator alone provides a nearly optimal predictor of its stable component. This is because the optimal predictor tends to put most of its weight on the corresponding indicator, i.e., the optimal predictor of stable SVA puts nearly all the weight on SVA and similarly for FFT and SSC. In simple terms, if states want to make inferences about teachers based on their contributions to the student learning as measured by the state test, they will gain little from using an optimal predictor rather than SVA.

Relatedly, optimal predictors of one stable component tend to be poor predictors of other stable components. In other words, there is an inherent tradeoff between forming a composite that is more aligned with one indicator but less aligned with others. For example, the correlation of the optimal predictor of FFT and the SVA stable components is less than 50 percent as large as the correlation between stable component and its optimal predictor. Similarly the best predictor of SSC also has correlation with the SVA stable component that is less the 50 percent as large as the optimal predictor except for elementary ELA, where the correlation reaches 72 percent of that of the best predictor. The results are similar for other stable components. An optimal predictor puts nearly all the weight on one indicator and that indicator which is not a good predictor of other

stable components because the stable components are not highly correlated.

Table 6: **Correlation Between Predictor and Alternative Composite Measures - SVA, FFT and SSC**

	SVA				**FFT**				**SSC**			
	Elementary		**Middle**		**Elementary**		**Middle**		**Elementary**		**Middle**	
	Math	**ELA**	**Math**	**ELA**	**Math**	**ELA**	**Math**	**ELA**	**Math**	**ELA**	**Math**	**ELA**
	Correlation with Best Predictor											
	0.72	0.61	0.92	0.69	0.68	0.65	0.84	0.84	0.72	0.67	0.95	0.94
	Correlation With Other Predictors Relative to the Best Predictor (%)											
Best predictor of SVA	100	100	100	100	37	46	50	25	45	72	48	40
Best predictor of FFT	37	46	50	26	100	100	100	100	58	59	61	54
Best predictor of SSC	45	72	48	40	58	59	61	54	100	100	100	100
SVA alone	99	93	100	98	28	24	45	14	32	42	42	21
FFT alone	25	29	36	20	98	97	97	99	40	39	43	39
SSC alone	33	53	45	40	45	41	57	51	98	96	100	100
FFT 50/SSC 50	36	53	48	37	90	89	92	91	87	87	86	85
SVA 50/FFT 25/SSC 25	90	97	90	91	70	67	79	62	71	79	74	64
SVA 25/FFT 50/SSC 25	64	73	70	62	94	94	96	91	74	78	74	69
Equal weights	76	87	79	76	83	81	87	78	82	88	81	77

Equal weighting or policy weighting (50-25-25) are also always suboptimal but they tend to be closer to the optimal than the other composites. For example, the correlation between stable SVA and equal and policy weighted composites is between 62 to 87 percent as large as it is with its optimal predictor. However, equal and policy weighted composites are also relatively highly correlated with FFT and SSC achieving over 80 percent of the correlation between each stable component and its optimal predictor. Overall, equal and policy weights result in a composite that is not optimal for any particular target, but close to optimal for many different targets.

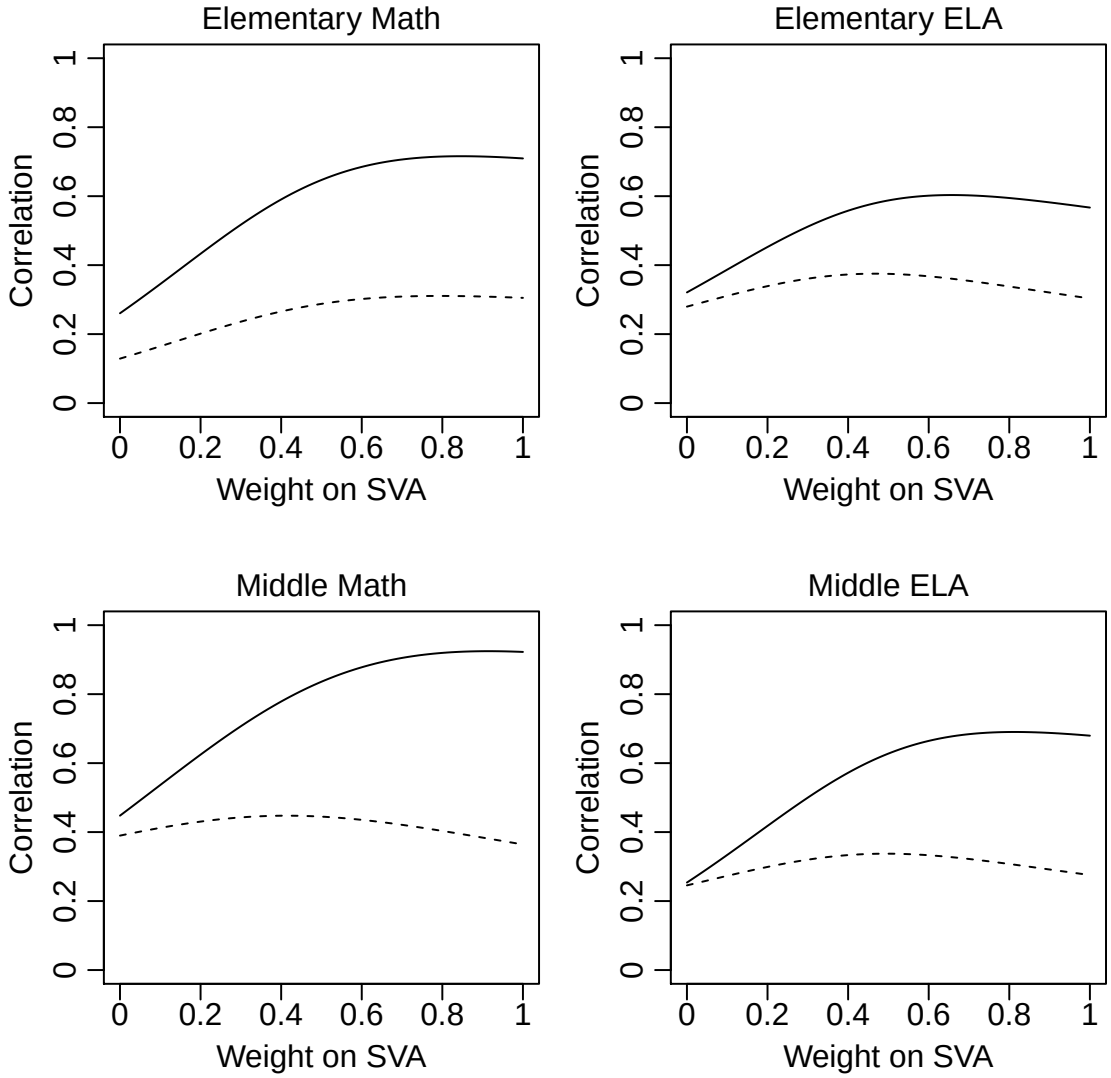
Table 7: **Correlation Between Predictor and Alternative Composite Measures - AVA**

	AVA			
	Elementary		**Middle**	
	Math	**ELA**	**Math**	**ELA**
Correlation with Best Predictor				
	0.31	0.38	0.45	0.35
Correlation With Other Predictors Relative to the Best Predictor (%)				
Best predictor of SVA	99	94	84	83
Best predictor of FFT	36	73	88	73
Best predictor of SSC	54	77	71	46
SVA alone	97	81	81	78
FFT alone	22	59	77	71
SSC alone	43	56	68	43
FFT 50/SSC 50	41	74	86	69
SVA 50/FFT 25/SSC 25	91	100	99	95
SVA 25/FFT 50/SSC 25	64	92	97	91
Equal weights	78	97	99	92

Table 7 repeats this exercise using AVA as the target criteria, but continuing to use SVA, FFT and SSC (and not AVA) to form the composite. This table illustrates the difficulty of predicting unmeasured targets such as performance on an alternative test. The optimal predictor for AVA has much lower correlation with stable AVA (.31-.45) than the optimal predictors of other stable components have with their targets (.61 or greater). This is because AVA is not in one of the indicators used in the composite. None of the indicators contain the unique component of AVA. The optimal predictor of AVA puts much of its weight on SVA because they share a common measurement mode, so the optimal predictor of AVA is highly correlated with the optimal predictor SVA, and both of these measures are correlated nearly as highly with AVA as the optimal (90 percent for the optimal predictor of SVA and 87 percent for SVA on average). However, because SVA is modestly

correlated with AVA, reducing the weight in SVA and increasing the weight on FFT and SSC does not have as dramatic of an impact on the correlation between the composite and stable AVA as it does between the composite and stable SVA. Equal and policy weights are even relatively efficient for AVA. The correlation between stable AVA and equal or policy weights is on average across subjects and grade levels about 91 percent as large as its correlation with its optimal predictor.

Figure 2: **Correlation of Composites with SVA and AVA** Each panel plots the correlation of a composite estimator of the form $w \times SVA + (1-w) \times (FFT + SSC)/2$ as a function of the weight w. The solid line is the correlation of the composite with SVA and the dashed line is the correlation of the composite with AVA.



Much of the recent policy debate has focused on how much weight should be placed on state value added in teacher evaluations. In Figure 2, we show how the weight placed on state value added affects the correlation of the composite with the stable component in SVA and AVA. For the figure, we consider composite estimator of the form $w \times SVA + (1-w) \times (FFT + SSC)/2$ for w from zero to one. At zero, we have a composite without SVA (FFT 50/SSC 50), at one we have SVA only, at $w = 1/3$ we have equal weights, and at $w = 1/2$ we have the SVA 50/FFT 25/SSC 25 policy weights. For each composite, we calculate the correlation between the composite and stable SVA (solid line) and stable AVA (dashed line) and we plot the correlation versus the weight on SVA.

There are three interesting features of Figure 2. First, the correlation between the composite and stable SVA is almost always greater than the correlation between the composite and stable AVA. Because AVA is not being included in the composite, and AVA has such a large unique component, no weighting scheme can produce a composite that is highly correlated with AVA. Second, the correlation between stable SVA and the composite is also more sensitive to weights and greatest when SVA receives a higher weight than the correlation between the composite and AVA. Thus, placing a low weight on SVA is costly if your target is to identify a teacher's impact on state test scores, but less costly in terms of identifying a teacher's impact on the alternative test. Finally, any composite that places a substantial weight on SVA (at least .4 in elementary, .6 in middle) will perform about as well as the best predictor in terms of alignment with a teacher's impact on state scores. For the teacher's impact on alternative test scores, an even broader range of weights result in composites that perform about as well as the best predictor.

When the target criteria is the teacher's impact on the state test, reducing the weight on SVA will quickly degrade the efficiency of the composite for predicting the target. However, the uniqueness of a teacher's impact on state tests (SVA) does not transfer to the teacher's impact on the alternative test (AVA) used in the MET project. If states are interested in generalizing to other tests rather than the specific state achievement tests, a lower weight on SVA might yield a better predictor of the quantity of interest. Our comparison of SVA and AVA in the MET project might overstate difference between value-added on different tests because the alternative test was different from the state test in terms of content, overlap, and format. However, other studies also suggest modest to weak correlation between value-added from alternative tests (Lockwood et al. ((2007)), Papay

((2010))).[19]

Moving forward states and districts might also consider observation protocols that are more specific, for example protocols developed to evaluate teaching of particular subject such as ELA. We repeated the analysis we conducted with observation scores from FFT with those from PLATO, which is an ELA specific protocol. The results are qualitatively nearly identitical as might be expected given the high correlation between the stable components of FFT and PLATO from Table 3. Figure C.1 in the appendix presents a graphical comparison the stability and the correlation with stable components for composites based on PLATO with those based on FFT.

5.5 Evaluating Predictions Under Alternative Data Collection Schemes

Table 8 presents the effect of different data collection schemes on the accuracy of the composite measure for predicting the stable component associated with each indicator. There are gains of up to 30 percent in the correlation with the stable component from increasing the number of years of data collection from one to three years of data, especially for elementary school teachers.

When predicting stable SVA, there is essentially no loss in correlation with the stable component from reducing the number students surveyed or the number of observations. There is a larger falloff in the correlation between the optimal predictor of FFT and stable FFT from cutting the number of observation by one fourth (from four to one). But there is essentially no loss from cutting the number of students surveyed. Similarly, there is a loss in the correlation between the best predictor of SSC and stable SSC from reducing the number of students surveyed of about five to eight percent, but there less than a one percent loss from cutting the number of observations. For AVA (Table 9), the correlation between the best prediction and stable AVA is not degraded by reducing the number of students surveyed but across subjects and grade levels the correlation declines by up to 7.5 percent from reducing the number of observations from four to one.

Cutting the number of surveys or observations also has a limited impact on the stability of the

[19] In addition to the correlations reported here, policymakers may be interested in teacher classification such as quartile rankings. Kane and Staiger (2012) show that the correlations reported in these tables are directly related to the rate of missclassification. In particular, suppose teachers are ranked into quartiles on the composite indicator. Then the correlation we report (between the composite and an underlying target) is approximately equal to the difference between teachers ranked in the top and bottom quartile in terms of the probability that their true performance (on the target) is above average. Thus, a correlation of 0.6 means that teachers ranked in the top quartile are about 60 percentage points more likely to be above average on the target than teachers in the bottom quartile of the composite.

optimal predictor of SVA because that composite puts little weight on the other indicators (not show in tables). It would reduce the reliability of the best predictor of AVA by up to 14 percent in ELA for both elementary and middle school teachers. Again most of the loss results from cutting the number of observations from four to one. For mathematics, the stability in the best predictor of AVA falls very little with a reduction in the number of students surveyed or observations conducted. The stability of the equal and policy weighted composites would decrease appreciably (over 25 percent in elementary ELA) if the number of surveys and observations were reduced. Cutting just the surveys would reduce the stability in these composites by about five percent or less; cutting the observations has the greater effect on the stability.

Table 8: **Alternative Data Collection Scenarios**

	SVA				**FFT**				**SSC**			
	Elementary		**Middle**		**Elementary**		**Middle**		**Elementary**		**Middle**	
	Math	**ELA**	**Math**	**ELA**	**Math**	**ELA**	**Math**	**ELA**	**Math**	**ELA**	**Math**	**ELA**
Correlation with Best Predictor Based on 1 Year of Data, 4 Videos Scored, All students surveyed												
	0.72	0.61	0.92	0.69	0.68	0.65	0.84	0.84	0.72	0.67	0.95	0.94
Correlation With Other Predictors Relative to the Best Predictor (%)												
2 years of data	114	120	104	116	116	117	107	107	115	116	102	103
3 years of data	121	130	105	123	125	126	111	111	121	125	103	104
Half students surveyed	100	99	100	100	100	100	100	100	92	93	96	95
Quarter videos scored	100	100	100	100	74	71	83	80	99	99	100	100

Table 9: **Alternative Data Collection Scenarios – AVA**

	Elementary		Middle	
	Math	**ELA**	**Math**	**ELA**
Correlation with Best Predictor Based on 1 Year of Data, 4 Videos Scored, All students surveyed				
	0.31	0.38	0.45	0.35
Correlation With Other Predictors Relative to the Best Predictor (%)				
2 years of data	114	117	104	112
3 years of data	121	126	105	118
Half students surveyed	99	99	100	100
Quarter videos scored	100	94	95	93

6 Discussion

States and districts are collecting multiple indicators of teaching and student outcomes to evaluate teacher effectiveness. There is limited empirical research to guide policymakers on how best to combine indicators to achieve specific goals or on the properties of such composite scores. Our analysis fills that gap, exploring the properties of different ways of combining multiple indicators using data from the Measures of Effective Teaching project. Our analysis looked at a range of indicators of effective teaching that are representative of the kinds of data currently being collected by many districts: student achievement growth, classroom observations, and student survey responses.

Individually, we found that all of these indicators captured a stable component of teacher performance. Under the data collection scenarios we considered, which were designed to represent what a typical district might consider, reliability was similar across all of the indicators, averaging about .4 in elementary grades and about .7 in middle grades. While the relatively low reliability of value added scores in elementary grades has been well documented, our results suggest that classroom observation and student survey indicators have similar levels of reliability, and reliability is higher in middle grades where a typical teacher has more sections.

We also found that the stable components of the indicators we considered were positively correlated, suggesting that all indicators capture a common factor related to effective teaching. While each indictor provided evidence about this common factor, the correlation among the stable compo-

nents was far from one, suggesting that each measure also captures some distinct unique dimension of effective teaching. Importantly, this implies that the weighting of individual indicators in a composite score is not purely a statistical question of placing more weight on the more reliable indicators, but also a policy decision regarding how much emphasis one wishes to place on the unique dimension captured by the stable component of each indicator: Composite scores which place heavy weight on any one indicator will tend to be more correlated with the unique dimension of that indicator and less correlated with the unique dimension captured by other indicators.

We evaluated a variety of ways of combining multiple indicators into composite scores. As a benchmark, we considered composite scores that were optimal predictors of the stable component of each indicator. These optimal predictors place heavy weight on the indicator corresponding to the stable component of interest, and are highly correlated (.61-.95) with the stable component they are intended to predict. From a practical perspective, if one is purely interested in the optimal predictor of the stable component of a given indicator, we find little gain from using a composite score over just using the indicator by itself (e.g. a composite score that places all of the weight on the indicator corresponding to the stable component of interest). Simply put, indicators of state value added are the best predictors of a teacher's impact on state test scores, classroom observation scores are the best predictors of a teacher's classroom practice, and student surveys are the best predictors of student perceptions of the teacher.

However, because the optimal predictors place heavy weight on the targeted indicator, we found they were not as good at predicting the stable component of other non-targeted indicators. Moreover, composites that place a very large share of the weight on a single indicator tend to have low stability relative to alternatives with more equal weights. Thus, placing a large weight on any one indicator yields a more unstable composite that predicts an idiosyncratic aspect of teaching but fails to provide an accurate prediction of other dimensions of effective teaching.

Composites with more equal weights are never the best predictor of the stable component of any particular indicator, but they tend to be more highly correlated on average with multiple different stable components. Composites with more equal weights also yield estimates that are more stable across years than composites with concentrated weights (so long as none of the observed indicators has very low reliability). Moreover, we found that composites with more equal weights

were also better predictors of teacher impact on an alternative low-stakes test that was not included in the composite. If districts value multiple dimensions of teaching - or value aspects of teaching that are not directly measured by any observed indictors, such as teacher contributions to student achievement broadly defined rather than restricted to the state accountability test - then setting weights near equal may be desirable. Overall, more equally weighted composites do almost as well as optimal predictors on multiple dimensions, and are more stable across years.

In general, our results were not particularly sensitive to details of the data collection such as how many years of data were used or the number of classroom observations or students surveyed. While additional data led to more accurate predictions, the differences between the scenarios we considered were not dramatic. For example, even an increase from one to three years of data only increased the correlation of composites with the stable components by at most 30 percent. The correlation of composites with the stable components of classroom observations and student surveys were somewhat sensitive to the number of observations and the number of student survey, respectively. If districts plan to use more equally weighted composites with interest in predicting teacher's classroom practice and student perceptions of the teacher, then more intensive data collection efforts for observations and surveys may be valuable. If the goal is to measure teacher contributions to the state accountability test then less intensive data collection efforts may be sufficient.

There are a number of important limitations of our analysis. We did not have measures of all outcomes that might be valued. In particular, our measures of "non-cognitive" skills were limited. An important unanswered question is whether teacher impacts on non-cognitive skills would also be predicted well by equally weighted composites, which one might expect if teacher impacts on non-cognitive skills shares the common component with the measures we used. In addition, our analysis was performed in a low-stakes environment (although some of the districts participating in our study were transitioning to high stakes for value added measures). Results may differ in high-stakes conditions where teachers would have more incentives to distort the individual indicators (particularly if composite scores placed heavy weight on any single indicator). In our exploration of equal and policy weights, we standardized the indicators to have variance one to remove differences in the scaling of the indicators. Districts using policy or equal weights would also need to remove scaling differences among their indicators. Finally, our results are conditional on the nature of the

data collection in the MET project, e.g., the video observers were well-trained and calibrated to master codes. In practice, the data collection might not yield equally reliable indicators. In such cases where indicators are measured with very low reliability our results suggest that those measures should receive a relatively low weight.

Creating weights for a composite estimator is a difficult task that involves trade-offs between competing priorities. As states and districts contemplate the weights for different indicators they must decide how much they value the unique dimensions captured by any particular indicator, knowing that those unique dimensions may be very specific to the indicator. More weight placed on any one indicator will identify teachers who perform better on that dimension but worse on other dimensions of teaching. More equally weighted composites scores, that average teacher performance across student achievement growth, classroom observations, and student survey responses, will not be optimal for targeting any particular dimension of effective teaching, but will be close to optimal across many dimensions and more stable across years.

References

Allen, M. J., & Yen, W. M. (2001). Introduction to Measurement Theory. Long Grove, IL: Waveland Press.

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

Brennan, R., & Johnson, E. (1995). Generalizability of Performance Assessments. Educational Measurement: Issues and Practice, 14(4), 9-12.

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)

Conover, W. (1999). Practical Nonparametric Statistics (Vol. 3). Wiley New York.

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.

Efron, B., & Tibshirani, R. (1994). An introduction to the bootstrap (Vol. 57). Chapman & Hall/CRC.

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

Gates, F. (2010). Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project (Tech. Rep.).

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.

Jacobs, R., Smith, P., & Goddard, M. (2004). Measuring Performance: An Examination of Composite Performance Indicators (Tech. Rep.).

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.

Kane, T., & Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation (Tech. Rep.). National Bureau of Economic Research Working Paper No. 14607.

Kane, T., Staiger, D., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., ... Parker, D. (2012). Gathering Feedback for Teaching: Combining High-quality Observations with Student Surveys and Achievement Gains (Tech. Rep.). (Seattle, WA: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project)

Kaplan, R. S., & Norton, D. P. (1992). The Balanced Scorecard - Measures That Drive Performance. Harvard Business Review, 70(1), 71-80.

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.

Lockwood, J., McCaffrey, D., Hamilton, L., Stecher, B., Li, V., & Martinez, F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. Journal of Educational Measurement, 44(1), 45-65.

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

Murias, P., de Miguel, J., & 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)

Papay, J. (2010). Estimates Across Outcome Measures Different Tests, Different Answers : The Stability of Teacher Value-Added. American Educational Research Journal, 48(1), 163-193.

Pinheiro, J., & Bates, D. (1996). Unconstrained Parametrizations for Variance-Covariance Matrices. Statistics and Computing, 6(3), 289–296.

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.

Robert, L. R. (2001). Generalizability Theory. New York:Springer-Verlag New York Inc.

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.

A Estimating $(\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{B})$

A.1 Maximum Likelihood Estimation

We used the section-level measures $\boldsymbol{Y}_{ij}$ and their associated sampling error covariance matrices $\boldsymbol{S}_{ij}$ to estimate the vector of means $\boldsymbol{\mu}$, the teacher-level covariance matrix $\boldsymbol{A}$ and the corresponding section-level covariance matrix $\boldsymbol{B}$ using maximum likelihood estimation (MLE). The estimation was conducted separately for elementary mathematics, elementary ELA, middle school mathematics and middle school ELA, because we wanted to allow the model parameters to vary across these factors and preliminary investigations suggested that some of them did.

As previously noted, in our estimation sample, teachers had at most 2 distinct sections, and a minority had only one section. We organize the 2 sets of component measures from each section as $(\boldsymbol{Y}_{i1}^{'}, \boldsymbol{Y}_{i2}^{'})^{'}$, a vector of length $2K$. We assume that this random vector is multivariate normal with mean vector $(\boldsymbol{\mu}^{'}, \boldsymbol{\mu}^{'})^{'}$ and covariance matrix $\boldsymbol{J}_2 \otimes \boldsymbol{A} + \text{diag}(\boldsymbol{B} + \boldsymbol{S}_{i1}, \boldsymbol{B} + \boldsymbol{S}_{i2})$ where $\boldsymbol{J}_2$ is a (2×2) matrix of 1s. This structure implies that measures are correlated across sections only through the shared teacher components. We assume the vectors of measures are independent across teachers so the joint likelihood function of $(\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{B})$ is a product of these individual multivariate normal likelihood functions. Some teachers had less than the full $2K$ measures either because they were observed on only a single section or because there was intermittent missingness on some components. Their resulting contributions to the likelihood function were based on only their observed components implying the assumption of missing at random (MAR).

Based on preliminary investigations examining each component measure in isolation, the video protocol measures (CLASS, FFT, MQI/PLATO) tended to have small or estimated zero variation at the section level. Only FFT had consistent evidence of non-zero section-level variation across grade levels and subjects. Models that permitted section-level variances and covariances for the other protocols led to lack of convergence of multivariate likelihood maximization. Therefore we conducted our estimation under the assumption that the variance and covariance elements of $\boldsymbol{B}$ corresponding to CLASS and MQI/PLATO were zero. Our results were not sensitive to an alternative model where we also forced the section-level variance and covariances of FFT to zero.

Maximization of the likelihood function was conducted in the R environment for statistical

computing. To improve numerical stability, the covariance matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ were parameterized using a log-Cholesky parametrization which is an unconstrained parameterization in which the parameters are the sub-diagonal elements, and the logs of the diagonal elements, of their respective Cholesky decompositions; i.e. $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^{'}$ for $\boldsymbol{L}$ a lower-triangular matrix (Pinheiro and Bates (1996)).

Calculation of the standard errors of the estimates of $(\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{B})$ using numerical evaluation of the second derivatives of the likelihood function evaluated at the MLE proved to be computationally unstable and time-consuming. In addition, some of the quantities for which we wanted standard errors, such as the values in Tables 1 and 3, were nonlinear function of the parameters. Therefore we obtained repeated samples of the estimates using nonparametric bootstrap sampling of teachers (Efron and Tibshirani (1994)), where a boostrap sample consisted of sampling teachers with replacement and taking all of their measures on both sections when applicable. We used 100 bootstrap replications for each of mathematics and ELA in elementary school, and 200 replications for each subject in middle school. These samples appeared to be large enough to provide sufficiently precise characterizations of estimation errors for quantities of interest.

We examined the robustness of the MLE estimates using an alternative Method of Moments (MOM) estimation strategy similar to that used in Kane and Staiger (2008). Briefly, the MOM strategy estimated the teacher-level covariance matrix $\boldsymbol{A}$ from the between-section covariance $Cov(\boldsymbol{Y}_{i1}, \boldsymbol{Y}_{i2})$, and estimated the section-level covariance matrix $\boldsymbol{B}$ by subtracting the teacher-level covariance matrix $\boldsymbol{A}$ and the average sampling error covariance matrix $\boldsymbol{S}_{ij}$ from the total covariance across the measures. These MOM estimates of $\boldsymbol{A}$ and $\boldsymbol{B}$, along with their standard errors, were very similar to the MLE estimates, suggesting that the multivariate normal assumption being used in the MLE is fairly innocuous in this application.

A.2 Estimation of Sampling Error Variance-Covariance Matrix

Our maximum likelihood estimates used estimates of the sampling error variance-covariance matrices $(\boldsymbol{S}_{ij})$ to allow for estimation of section-level variance-covariance matrix. We used disaggregated data to estimate the elements of $\boldsymbol{S}_{ij}$. The diagonal elements of $\boldsymbol{S}_{ij}$ corresponding to student-based measures (SVA, AVA, SSC, EFF and HIC) equal the variance of the student-level errors for the indicator, s_{kk}, divided by the number of students used in calculating the measure for the section,

n_{ijk}: $\boldsymbol{S}_{ij}[kk] = s_{kk}/n_{ijk}$ The off-diagonal elements equal the covariance among the errors from pairs of measures, $s_{kk'}$, times $n_{ijkk'}/(n_{ijk}n_{ijk'})$, where $n_{ijkk'}$ equals the numbers students in the sections who contributed to both indicator k and k': $\boldsymbol{S}_{ij}[kk'] = s_{kk'}n_{ijkk'}/(n_{ijk}n_{ijk'})$.

We used the pooled within section squared error of the residuals from our first stage regression for value-added to estimate an error variance components s_{kk} for SVA and AVA. We used the covariance of the residuals to estimate the covariance among the errors in the two value-added scores. We used the pooled within section variance and covariance of the survey measures to estimate the variance and covariance components for those measures and the covariance between these measures and the value-added errors.

The ζ_{ijk} for observation-based indicators depended on the lessons observed and the raters conducting the observations. We assumed these errors were independent of the individual student level measures, although we did allow observation-based and other indicators to have correlated section-level errors determined by estimates of the elements of $\boldsymbol{B}$. To estimate the variance of ζ_{ijk} for an observation-based indicator we fit a generalizability study, G-study, (Brennen (1995), Kane et al. (2012)) to the individual ratings of each lesson. We used the variance components from the G-study along with the number of lessons and ratings per lesson to calculate the corresponding diagonal element of $\boldsymbol{S}_{ij}$. To estimate covariances among protocols we averaged scores from the same video for each protocol and then estimated the within lesson covariance among of scores for pairs of protocols.

A.3 Estimating Optimal Weights for Individual Teachers

If states or districts want to use optimal weights for their composite, they can use the following process.

1. Collect from multiple sections for each teacher data on the indicators for the composite and the target criterion (districts must have data on the target criterion to estimate optimal weights).
2. Using disaggregated data estimate the variance and covariance among the student-level and lesson-level error in scores from the multiple measures and calculate $\boldsymbol{S}_{ij}$ for every section.
3. Use the maximum likelihood estimation methods describe above to estimate $\boldsymbol{A}$ and $\boldsymbol{B}$.
4. Combine section-level the section level data into teacher level indicators.

5. Calculate $\boldsymbol{B}_i$ and $\boldsymbol{S}_i$ using the maximum likelihood estimates and the numbers of students and sections that provide data for the teacher.

6. Using Equation 3 and the estimates of $\boldsymbol{A}$ and $\boldsymbol{B}_i$ and $\boldsymbol{S}_i$ for the teacher determine the optimal weights for the teacher.

Because data from different teachers will be from varying numbers of students and sections and possibly different numbers of observations or other measurements, $\boldsymbol{B}_i$ and $\boldsymbol{S}_i$ will differ across teachers and the weights applied to the various indicators may differ across teachers too.

An alternative approach to estimating optimal weights is to collect data on the indicators and the target criterion for two school years, year 1 and year 2. Using linear regression, model the target criterion from year 2 as a function of the indicators from year 1. The estimated coefficients are the "optimal" weights. These estimates of the optimal weights will be constant across teachers and therefore biased when $\boldsymbol{B}_i$ and $\boldsymbol{S}_i$ vary across teachers because the true optimal weights must differ across teachers if $\boldsymbol{B}_i$ and $\boldsymbol{S}_i$ do. The estimated optimal weights will not result in the truly optimal predictions of the target criterion. However, in limited simulations and empirical investigations we find that the linear regression based estimates of the optimal weights yielded estimates that were nearly optimal on average across teachers although they could be inefficient for some teachers with very small or large classes.

B Formula for Fit Statistics

We consider a composite measure of the form $\gamma' Y_i$, where γ is a vector of weights (optimal or other) for the indicators Y_i. We use to metrics to evaluate this composite: the correlation with the target criterion and the stability of the composite across years. We assume that the target criterion η_i has variance ν^2.

B.1 Correlation

The correlation of the composite $\boldsymbol{\gamma}'\boldsymbol{Y}_i$ with η_i is

$$R_{\gamma} = \frac{\boldsymbol{\gamma}^{'}\boldsymbol{c}}{\sqrt{\boldsymbol{\gamma}'(\boldsymbol{A}+\boldsymbol{E_i})\boldsymbol{\gamma}\boldsymbol{\nu^2}}}, \tag{4}$$

where $\boldsymbol{c}$ is the vector of covariances between η_i and $\boldsymbol{Y}_i$. If $\boldsymbol{\gamma}$ are the optimal weight ($\boldsymbol{\gamma} = \boldsymbol{c}'(\boldsymbol{A}+\boldsymbol{E}_i)^{-1}$) then

$$R_{opt} = \sqrt{\frac{\boldsymbol{c}'(\boldsymbol{A}+\boldsymbol{E}_i)^{-1}\boldsymbol{c}}{\nu^2}}. \tag{5}$$

B.2 Stability

We assume the stable components of the indicators are constant across year and that all measurement and section-level error are independent across year and with equal variance in both years. Given these assumptions, the covariance between measures from adjacent years equals

$$cov(\boldsymbol{\gamma}'\boldsymbol{y}_{it}, \boldsymbol{\gamma}'\boldsymbol{y}_{it+1}) = \boldsymbol{\gamma}'\boldsymbol{A}\boldsymbol{\gamma}. \tag{6}$$

Consequently, the correlation equals

$$cor(\boldsymbol{\gamma}'\boldsymbol{y}_{it}, \boldsymbol{\gamma}'\boldsymbol{y}_{it+1}) = \frac{\boldsymbol{\gamma}'\boldsymbol{A}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'(\boldsymbol{A}+\boldsymbol{E}_i)\boldsymbol{\gamma}}. \tag{7}$$

We use our estimates of $\boldsymbol{A}$ and $\boldsymbol{V}$ for the various data collection and multiple values for $\boldsymbol{\gamma}$ to assess the stability of the composites under different data collection scenarios.

C Comparison of Composites Based on PLATO with those Based on FFT

We reran our analysis in Section 5 for ELA teachers replacing FFT with PLATO. Figure C.1 presents the stability of the resulting predictors and their correlation with stable component. It also presents

the corresponding values for predictors using FFT.

Figure C.1: **Comparison of Predictions with PLATO vs. FFT.** The left panel is for elementary teachers. The right panel is for middle school teachers. In each panel the cells in the first column present the stability of predictions. The cells in the remaining columns present the correlation between the predictions and stable components. In each cell, the light gray bar is for the prediction using PLATO and the dark gray bar is for the prediction using FFT. In the column labeled "Observation" for PLATO the correlation is between predictions and the stable component of PLATO; for FFT the correlation is between predictions and the stable component of FFT. In the row labeled "Best Obs", the predictions are for predicting the stable component of PLATO, for FFT the predictions are for the stable component of FFT.