Exploring Explanations for the "Weak" Relationship Between Value Added and Observation-Based Measures of Teacher Performance

Mark Chin and Dan Goldhaber

1. Introduction

Since 2009, 49 states and the District of Columbia have changed their teacher evaluation systems in response to federal incentives, such as flexibility waivers to No Child Left Behind and Race to the Top grants.[1] In many cases teacher evaluation reforms have included the use of student growth, or "value-added", measures of teacher performance. These measures of teachers' contributions to student performance on standardized tests represent a relatively new way to assess practicing teachers, though value-added models have been employed as an analytic tool for decades by researchers (e.g., Hanushek, 1971; Murnane, 1981). Value-added measures are also controversial (Baker et al., 2010; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012) and can only be used to assess teachers in tested grades and subjects, who represent less than 33 percent of the teacher workforce (Papay, 2012). Not surprisingly, given their history as an evaluation tool that can be used to assess all teachers, virtually all states also include observations of teachers' classroom practice as a component in a summative evaluation (Doherty & Jacobs, 2013).

It is unclear what the relationship ought to be between value-added and observational measures, but the relationship is often characterized as being "modest" or "weak" (e.g. Harris, 2012). Moreover, some judge the relationship between these measures (described more extensively below) to be problematic for use by policymakers who might wish to use value added and observations together to identify effective or ineffective teachers. Audrey Amrein-Beardsley (2014), for instance, notes that "value-added scores do not align well with observational scores, as they should if *both* measures were to be appropriate[ly] capturing the 'teacher effectiveness' construct". Notwithstanding the characterization of the relationship between value-added and observational measures, several scenarios exist that result in a weak

[1] See Minnici, 2014.

correlation; not all of them suggest that the two measures capture different teacher effectiveness constructs. Variation in the multidimensionality, validity,[2] and reliability of value added and observations distinguish these scenarios from one another.

Few studies have investigated the scenarios that might explain attenuated correlations between value-added and observational measures, or have suggested which are unlikely given observed correlations in prior research. Our paper explicitly illustrates these different scenarios, and uses simulated data to formally investigate the extent to which one or another explanation is likely to explain weak correlations between the measures. We explore the levels of correlation between value-added and observation scores after varying two broad factors. First, we adjust the correlation of each teacher's score on an underlying dimension of "teacher quality" to its two different proxy measures: error-free value added and error-free observational measures of teacher practice. This adjustment allows us to investigate the effect of changes in the validity of these measures. Second, we add error to these measures to create simulated outcomes (i.e., "student test performance" or "lesson performance"), and vary the number of outcomes used to estimate measure scores. This adjustment allows us to investigate the effect of changes to measure reliability. With the results from our simulations, we attempt to answer the following research question: What is the magnitude of the correlation between value-added and observation scores, given different levels of validity and reliability for each measure of teacher quality?

In what follows, we recount the historic use of value-added and observational measures in teacher evaluation systems, the research on their relationship, and the factors that impact this

[2] We discuss two types of measure validity in our paper. The first type refers to the extent to which value added and observations serve as good proxies for some desirable underlying dimension or dimensions of teacher quality. The second type refers to the extent to which the performance of a teacher's students on tests, or the performance of a teacher during observed lessons, reflect his or her true value-added or observation scores, respectively (also referred to as "systematic error", see McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). We use the term "validity" to represent the first type, unless otherwise specified.

relationship (Section 2). We then (in Section 3) discuss the design of our simulation and the parameters we vary to reflect these key factors. After describing our process for creating the simulated data and method of analysis, we discuss the simulations' results (Section 4). Finally (in Section 5), we discuss the implications for researchers and practitioners and offer some concluding thoughts.

2. Value-Added and Observational Measures of Teacher Quality and Their Relationship

Value-added methods have long been used as a means of assessing both educational productivity and the effects of specific schooling inputs (e.g., Hanushek, 1971; Murnane, 1981). They have also been used to assess the implications of differences amongst individual teachers and the extent to which individual teachers explain the variation in student test performance (e.g., Goldhaber, Brewer, & Anderson, 1999; Hanushek, 1992; Nye, Konstantopoulos, & Hedges, 2004). Though a few states and districts began using value-added and other related test-based measures of teacher quality in the late 1990s (Sanders & Horn, 1998), it is only in recent years that the use of value-added measures has proliferated across the nation. This proliferation has engendered debates amongst researchers and policymakers about whether value added is a fair measure of teachers' contributions in the classroom, and, relatedly, how its use will affect teachers and students.

Value added has been linked to long-term student outcomes (Chetty, Friedman, & Rockoff, 2014b) and been shown to be unbiased in some experimental and quasi-experimental settings (Bacher-Hicks, Chin, Kane, & Staiger, in preparation; Chetty, Friedman, & Rockoff, 2014a; Kane & Staiger, 2008; Kane, McCaffrey, Miller, & Staiger, 2013). Yet questions remain about the extent to which value added may be used to obtain unbiased estimates of teacher performance (Rothstein, 2008, 2014), and, even if the measures are unbiased, whether they are

stable enough from year to year to use,[3] or would have negative ramifications for teacher behavior (Baker et al., 2010; Darling-Hammond et al., 2012).

Scholars have similarly investigated the quality of teachers through their practices in the classroom for decades (Brophy & Good, 1986). Compared to value-added measures, classroom observations of teaching have longer played a role in evaluation systems, yet have not faced the same level of academic scrutiny as value added (Corcoran & Goldhaber, 2013).[4] Recent findings, however, have found that the traditional observation systems used in some states and districts failed to meaningfully differentiate teachers (Weisberg, Sexton, Mulhern, & Keeling, 2009). Revisions to preexisting observation systems have led some locales to adopt observation protocols developed by the academic community, such as the Danielson Group's *Framework for Teaching* (Herlihy et al., 2014). These protocols, which are also widely used in research projects, identify key classroom practices that, in theory, should be important for student learning, and also standardize how teachers are evaluated on these practices.

The relationship between value added and observations

A number of the recently implemented educator evaluation reforms include the use of multiple measures of teacher quality, and many states and districts use both value-added and observational measures when assessing teachers' performance (Herlihy et al., 2014). Not surprisingly, there is a growing research base that explores the extent to which these measures are related to one another. For example, the *Measures of Effective Teaching* (MET) project, a large scale study of teacher quality, explored the relationship of teacher value added and observations and found correlations between the two measures ranging from 0.12 to 0.34,

[3] See Goldhaber and Hansen (2013) and McCaffrey, Sass, Lockwood, and Mihaly (2009) for estimates of the stability of value added.
[4] See Cohen and Goldhaber (2015) for a review of this role and a comparison of what we know about the properties of observations and value added.

depending on the observation protocol (Kane & Staiger, 2012). With some exceptions (e.g., Schachter & Thum, 2004), most other recent studies have replicated this pattern of a weak or moderately weak relationship when analyzing similar observation protocols (e.g., Bell et al., 2012; Grossman, Loeb, Cohen, & Wyckoff, 2013; Hill, Kapitula, & Umland, 2011; Kane, Taylor, Tyler, & Wooten, 2011). These findings contradict what many scholars and practitioners might expect. Theory and intuition suggests that strong instructional practices by teachers should lead to improvements in student test performance. In this paradigm, value-added and observation scores should be highly correlated.

Furthermore, states and districts have practical reasons to be concerned about the weak relationships observed in extant literature. A weak relationship may indicate that one or both are not valid measures of some dimension of teacher quality. It also sends contradicting signals to practitioners about their strengths and weaknesses, which in turn may inhibit the improvement of teachers' practice (Polikoff, 2014). Finally, it could serve to undermine the trust in teacher evaluation systems, making it more politically difficult to use evaluations to inform key personnel decisions such as compensation or tenure (Herlihy et al., 2014).

Explanations for the weak relationship between value added and observations

There exist at least three scenarios that result in weak correlations between value added and observations. The first is that one or both measures could provide unreliable estimates of one or more dimensions of teacher quality, due to sampling error. The second is that teacher quality may be multidimensional, and the measures provide reliable estimates of different dimensions of teacher quality. And the third is that one or more of the measures may be invalid, in the sense that the measure does not provide a reliable estimate of any dimension of teacher quality. We provide simple illustrations of these scenarios in **Figure 1**.

[Figure 1 about here]

In Panels A and B of the figure, we depict underlying dimensions of teacher quality (TQ) with the bullseyes in the targets. In practice, we use value added and observations to serve as proxy measures for each teacher's quality, which we never observe. We also never observe each teacher's true, error-free value-added or observation score. Instead, we estimate value-added and observation scores from two different observed outcomes, represented in the figure: student test performance (v) and performance on lessons (o), respectively. The clouds around each set of outcomes show the distribution of the data points used to estimate each measure, with a darker color representing estimates based on the aggregation of information from each measure (e.g. from multiple student test results, or multiple observed lessons). The dashed, two-headed arrow represents the distance or correlation between the two different measures of teacher quality; a shorter arrow indicates that the two measures align more closely. Moving from the left target to the right in either Panel A or B of Figure 1, the amount of information for each measure of teacher quality increases (e.g. through more having more students' test results or observing teachers' lessons more often), increasing the reliability of each measure.

The leftmost illustration in Panel A depicts the first scenario for weak correlations, where both measures would serve as valid proxies for the same dimension of teacher quality, but are estimated unreliably. Value added and observations could be estimated unreliably due to factors such as observing a teacher on a particularly good or bad day, or analyzing the test results of students who by chance perform well or poorly on a test; either would add sampling error to scores. To counteract sampling error in value added, many research projects will estimate teachers' value added using Empirical Bayes estimators, which shrink scores that are estimated

less reliably (e.g., estimated from the test performance of fewer students) toward the mean (e.g., Kane & Staiger, 2008; Sanders & Horn, 1994).[5] Another way to counteract sampling error is to estimate value added and observations with as many data points as possible. For example, the stability of value-added measures, moderate when estimated from a single year of student test performance data (McCaffrey et al., 2009), improves when using multiple years of data (Goldhaber & Hansen, 2013). Though states and districts need to consider the financial and temporal burdens associated with reducing sampling error in value added and observations by increasing data points, improving measure reliability would disattenuate the relationship between both.

In research and practice, teacher value added and performance on observations are often treated as measures of the same underlying construct—the scenario depicted in Panel A. However, there are reasons to believe that they are not, and that Panel B of Figure 1 depicts a more accurate representation of reality. Panel B of the figure illustrates a case where there are two dimensions of teacher quality (TQ1 and TQ2) and each measure of quality is a reliable estimate of only one of the dimensions. For example, one dimension might capture the degree to which teachers contribute to student knowledge, and a second dimension might be the extent to which teachers contribute to students' ability to interact productively with one another. These dimensions of teacher quality may or may not be closely related, and the correlation between the measures of teacher quality may or may not increase as the reliability of each measure increases. In the example depicted by Panel B, the correlation between the measures decreases (i.e., the arrows become longer) as each measure of teacher quality becomes more reliable, moving from

[5] In theory, the same adjustment for reliability can apply for observations as well. In practice, however, little research appears to use Empirical Bayes estimators to adjust scores for differences in the number of lessons observed. However, it is not clear that such estimates provide the best indicator of teacher effectiveness (see, for instance, Mehta, 2015).

the left target to the right. Thus, the rightmost target in Panel B illustrates the second scenario for a low correlation between the measures: that each measure provides a reliable estimate of different dimensions of teacher quality.

Some empirical evidence substantiates this second explanation for weak correlations. For instance, prior research suggests that measures of teacher contributions to the performance of students on different tests may themselves capture divergent dimensions of teacher quality. The most obvious example of this divergence emerges when comparing teachers' value added in different subjects; for example, one might not expect a teacher's contributions to performance on a mathematics exam to be measuring the same type of quality as his or her contributions to performance on a reading exam (Fox, forthcoming; Gershenson, forthcoming; Goldhaber, Cowan, & Walch, 2013; Rockoff, 2004).

Even across tests of the same subject, value-added scores are somewhat sensitive to the specific tested items. Sass (2008) compares teachers' contributions to performance on a high-stakes test versus a low-stakes test in the same subject and finds a correlation of 0.48. Papay (2011) finds that the correlations between teachers' value added on three different reading tests range from 0.15 to 0.58, depending on the compared tests and included value-added model controls. He also replicates the findings of Lockwood and colleagues (2007), who found weak to moderately weak relationships between teacher contributions to different subscales within the same mathematics test. These results indicate that value-added measures themselves may not be measuring the same construct of teacher quality, even within a subject. Instead, teachers likely vary in their ability to improve performance on questions that cover different subtopics, require different skills, or even have different implications for accountability.

Arguments of multidimensionality also apply within observational measures (McClellan, Donoghue, & Park, 2013). Theoretically, observation protocols used in research and practice, such as the Danielson *Framework For Teaching*, the *Protocol for Language Arts Teaching Observations* (Grossman, Cohen, Ronfeldt, & Brown, 2014), the *Mathematical Quality of Instruction* (Hill et al., 2011), or the *Classroom Assessment Scoring System* (Bell et al., 2012) all broadly capture a teacher's effectiveness in delivering quality instruction. Each, however, evaluates teachers on different subjects and classroom practices.

Because multidimensionality may exist within types of measures of teacher quality, it thus follows that multidimensionality, and subsequently, weak relationships, across value added and observations would be expected. [6] This expectation is further supported by the fact that, by design, both measures directly assess teacher impacts on different types of outcomes for different populations. For example, the former broadly assesses student test performance, while the latter broadly assesses teacher observational protocol performance. Research has indicated teacher quality to be multidimensional even when measuring outcomes, other than test performance, of just students (Gershenson, forthcoming; Jackson, 2012; Jennings & DiPrete, 2010). For example, Gershenson uses student administrative data from North Carolina and finds insignificant and even negative correlations when comparing teacher rankings in terms of effects on student absences and on effects on student test performance.

The last scenario for weak correlations between value-added and observations relates to the validity of the measures. This scenario can also be depicted by Panel B of Figure 1, where, instead of measuring a dimension of teacher quality, value added or observations provide a reliable estimate for a completely unrelated construct. An observation protocol, for instance,

[6] In cases where the assessed skills of student tests and observation protocols align, researchers have indeed found stronger relationships between teachers' contributions to student test performance and observation scores (Grossman et al., 2014; Lynch, Chin, & Blazar, submitted).

might preference classroom practices that are thought to be related to some dimension of quality teaching when in fact they are not.[7]

One final property of value-added and observational measures, unrelated to their validity as teacher quality proxies or their reliability due to sampling error, affects the correlations between scores resulting from *any* given scenario: the validity of the observed outcomes used to estimate either measure. As noted earlier, like underlying teacher quality, we never directly observe a teacher's true ability to impact student test performance or deliver quality instruction. These measures are instead estimated using actual student test scores or observation scores for individual lessons. Thus, estimates of value added or observations may themselves serve as invalid proxies for true value-added and observation scores if other factors besides teachers largely influence these scores. Under this scenario, we would expect attenuated correlations, as well.

Many observed and unobserved factors, potentially unrelated to teachers, influence student test performance. For example, research shows that the choice of student, classroom, and school-level controls in models used to estimate value added influences the amount of variance in the outcome associated with teachers. [8] Adding controls, which decrease teacher-level variance, may result in an underestimation of teacher contributions if, for example, controlling for factors such as the effect of students' peers removes a component of true teacher quality.[9] Conversely, many opponents of value-added measures argue that estimation models do not account for enough. They note that these scores may also fail to take into consideration the effect

[7] In practice, the two scenarios of Panel B cannot be extricated from one another, because the underlying constructs are never actually observed.
[8] For an overview of model specifications used in research and practice, and the implications of these specifications, see Goldhaber and Theobald (2013).
[9] Using experimental data, Kane and colleagues (2013) indicate that the component of teachers' effects on student test performance removed by adding controls for peer effects in fact removes a component of the teacher's true quality. They do not find this to be the case for controls for student prior achievement.

of learning gains or losses outside of the classroom (Papay, 2011) and the systematic sorting of students to teachers (Rothstein, 2009). Despite these questions of model specification, empirical evidence suggests that the magnitude of teacher effects on student test performance to be fairly small, yet stable across studies; differences between teachers generally account for between one to 10% of the variance in performance on reading and mathematics tests (Hanushek & Rivkin, 2010).

Similarly, a variety of observed and unobserved sources contribute to a teacher's ability to deliver effective instruction in an individual lesson observation. For example, observation scores may be subject to differences between raters performing teacher observations; findings from various generalizability studies (see Shavelson & Webb, 1991) on aforementioned observation protocols support this possibility (e.g, Casabianca, Lockwood, & McCaffrey, 2014; Casabianca et al., 2013; Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013; Mashburn, Meyer, Allen, & Pianta, 2013). This rater component in scores has been found when investigating the effect of external versus traditional (i.e., within-school administrators) raters (Ho & Kane, 2013); ratings done live versus using video-recorded lessons (Casabianca et al., 2013); and ratings done on observations of varying length (Mashburn et al., 2013). Many factors can result in differences between raters. For example, the quality of certain instructional behaviors may be more subjective, or raters may possess varying amounts of specialized content knowledge necessary to assess content-specific teacher practices. Even knowledgeable raters, however, have personal biases and varying levels of experience with scoring procedures of observation protocols (Casabianca et al., 2013), which may lead to assessments of individual lessons that fail to capture teachers' true observation score.[10] Compared to value added,

[10] The effect of the student composition of a teacher's classroom on his or her performance on observational protocols has been much less explored by research (see Cohen & Goldhaber, 2015). Initial evidence from an

however, estimates of the teacher-level variance components for estimated observation scores are far less stable across studies. For example, the MET project reported components for observation scores using different protocols from as low as six percent to as high as 37 percent. In spite of this wider range, however, overall differences between teachers generally appear to have a larger effect on individual lesson performance compared to their effect on individual student test performance.

Though existing research theoretically supports that weak relationships between value-added and observational measures may result from any of the three outlined explanations, little work has investigated the scenarios concurrently to compare their likelihoods in representing reality. Each explanation bears implications in research and practice on how scores might be estimated (e.g., from more data points) or used and interpreted as measures of teacher quality (e.g., as measuring distinct dimensions of quality, or not measuring a dimension at all). Thus, we systematically explore each of the scenarios using simulated data.

3. Simulation Design and Methods

Design

By identifying the explanations in each scenario for why the relationship between measures of teacher quality may be weak, we are able to investigate the effect on correlations after varying these factors in a simulated study. In **Table 1**, below, we describe the parameters that we vary in our simulation to specifically reflect changes in the degree of measure validity and reliability.

[Table 1 about here]

---

analysis of observational protocols, similar to the analysis on value-added by Kane and colleagues (2013), indicates that adding controls for peer effects removes a component of the teacher's true quality, as judged by performance on the MQI observation protocol, but not necessarily for the CLASS (Bacher-Hicks et al., in preparation).

The first two parameters we adjust are the correlation of each teacher k's true score on an underlying dimension "teacher quality" (TQ_k) to two different error-free proxy measures: value added (VA_k) and observations (OBS_k). We have chosen to model only a single dimension of teacher quality even though some prior work has suggested teacher quality to be multidimensional. However, as noted above no distinction exits between the scenario where one or both *measures* of teacher quality are invalid from the scenario where the measures point to different dimensions of teacher quality. For parsimony's sake we henceforth treat the discussion as if the variation in the correlations between our measures and a single dimension of teacher quality speak to the validity of each measure. We discuss the implications of our simulation in the conclusion considering the alternative possibility.

We choose correlation levels for teacher quality error-free value added ($\rho_{\mathrm{VA,TQ}}$) and teacher quality with error-free observations ($\rho_{\mathrm{OBS,TQ}}$) to range from 0.05 to 0.95, at intervals of 0.05.[11] The other two parameters we adjust are the number of students or lessons used to estimate each teacher k's value added ($\mathrm{VA'}_k$) and observations (OBS'_k) with error, respectively. These parameters allow us to investigate the effect of measure reliability on correlations. We vary the number of students ($\mathrm{N}_{\mathrm{VA}'}$) that we use to estimate value added to reflect a reasonable range of student data points that might be used to construct an upper-elementary school teacher's value added; we adjust the parameter to range from 15 to 35 students. We similarly vary the number of lessons ($\mathrm{N}_{\mathrm{OBS}'}$) that we use to estimate observation scores to reflect a reasonable range for number of lessons that a teacher might be observed on in a given year for evaluative purposes; we adjust the parameter to range from one to six lessons. Because increasing the

[11] Because we never observe the true, underlying dimension of teacher quality that value added and observations serve as proxies for, the actual correlation between these measures to teacher quality is unknown. Thus, we test a range of values when investigating validity.

number of data points used in estimation reduces the amount of sampling error in scores, we expect stronger correlations between value-added and observation scores when the number of students or lessons is higher.

We fix three different parameter to be constant for all of our simulations. First, we simulate value-added and observational measures for 400 teachers. [12] Second, for estimating value added, we fix the percentage of variance in student test performance that differences between teachers account for to be 7.5%.[13] Finally, for estimating observation scores, we fix the percentage of variance in lesson performance that differences between teachers account for to be 30%.[14]

Methods

From the parameters outlined above, there are a total of 45,486 potential unique combinations of value-added and observation scores, estimated with error, that we correlate in our simulations. For each of these combinations, we run 100 simulations. In each of the simulations, we first randomly generate three scores for each of 400 "teachers", using the fixed values for percent variance in outcomes (i.e., student test performance or observed lesson performance) at the teacher-level to define the distribution of scores:

$\mathrm{TQ}_k \sim N(0,1)$, $\mathrm{VA}_k \sim N\left(0, \sqrt{0.075}\right)$, and $\mathrm{OBS}_k \sim N(0, \sqrt{0.30})$.

To estimate $\mathrm{VA'}_k$, we first create for each teacher $\mathrm{N_{VA'}}$ "students". Each student has a "test performance" score, generated from the following equation:

[12] Recent value-added research has utilized a range of values for number of teachers when simulating data, from 120 teachers (Guarino, Reckase, Stacy, & Wooldridge, 2015) to 600 teachers (Goldhaber & Chaplin, 2015). We decided to simulate data for 400 teachers, reflecting a number used by Glazerman and colleagues (2011), which also falls within the range used in other studies.

[13] This percentage of variance of student test performance at the teacher-level is similar to the one observed in Goldhaber et al. (1999), and also results in a teacher effect size similar to those reported in Hanushek and Rivkin (2010) across studies investigating student mathematics test performance.

[14] This approximates the average of the teacher-level variance components for overall observation protocol scores seen in the MET project (Kane & Staiger, 2012).

(1) $\mathrm{STU}_{jk} = \mathrm{VA}_k + \varepsilon_{jk}$

Each student j's test performance, $\mathrm{STU}_{jk} \sim N(0,1)$, is a function of his or her teacher k's error-free value added, VA_k, the percent variance in student test performance due to factors besides differences between teachers, and sampling error, $\varepsilon_{jk} \sim N(0, \sqrt{1 - 0.075})$. We then take the average test performance of all teacher k's students to arrive at that teacher's value added estimated with error, $\mathrm{VA'}_k$.

We follow a similar procedure to estimate $\mathrm{OBS'}_k$. We create for each teacher $\mathrm{N}_{\mathrm{OBS'}}$ "lessons". Each lesson has a "performance" score, generated from the following equation:

(2) $\mathrm{LES}_{lk} = \mathrm{VA}_k + \varepsilon_{lk}$

Each lesson l's performance, $\mathrm{LES}_{lk} \sim N(0,1)$, is a function of the teacher k's error-free observation score, OBS_k the percent variance in observed lesson performance due to factors besides differences between teachers, and sampling error, $\varepsilon_{lk} \sim N(0, \sqrt{1 - 0.30})$. We then take the average performance of all teacher k's lessons to arrive at that teacher's observation score estimated with error, $\mathrm{OBS'}_k$. [15]

We use correlations between this score with estimated value added ($\rho_{\mathrm{VAM'},\mathrm{OBS'}}$) to help answer the following research question: What is the magnitude of the correlation between value-added and observation scores, given different levels of validity and reliability for each measure of teacher quality?

4. Simulation Results

[15] There are many approaches to estimating value-added or observation scores in research and practice. For example, mixed modeling with fixed effects or random effects may be used (see Sanders & Horn, 1994). Random effects, in particular, can be estimated with Empirical Bayes, which would shrink scores to the mean, due to lower reliabilities resulting from fewer data points used in estimation. Because we estimate each teacher's scores from the same number of data points, however, we do not use random effects. Furthermore, modeling scores with fixed effects and random effects change the magnitude of scores. Because we use simple correlational analysis, however, this adjustment is also unnecessary. Correlations between simulated value added and observations do not change regardless of which method of estimation we use.

We begin by reporting the overall patterns in correlations between value added and observations associated with various levels of validity and reliability of each. **Figure 2** below shows how correlations, averaged across the 100 simulations for each combination of parameters, change as the number of data points increase and the underlying correlation between both proxy measures with teacher quality increase.

[Figure 2 about here]

The simulated results illustrate the fact that as the correlation between underlying value added to teacher quality (depicted on the subgraph y-axes) increases, or between underlying observations to teacher quality (depicted on the subgraph x-axes) increases, the average correlation between the value added and observations measures also increase. This is depicted by movement, within subgraphs of Figure 2, from "red" regions (i.e., correlations less than 0.15) to "yellow" (i.e., correlations between 0.15 and 0.30), "green" (i.e., correlations between 0.30 and 0.45), and "blue" (i.e., correlations greater than 0.45) regions, as one moves up the y-axes or right on the x-axes.

As the number of students informing the value-added measure or lessons informing observation measure increases (across subgraphs along the y- and x-axis, respectively), the reliability of each of the measures increases, which in turn increases the correlations of the two measures with each other. However, another pattern also emerges when investigating the effects of increased measure reliability: the rate at which average correlations between measures improve as you estimate value added and observations with more data points is far greater for additional lessons than additional "tested" students. For example, the proportion of average correlations that are greater than 0.45 is zero when estimating value added from 15 students and observation scores from one lesson. Estimating observations from at least two lessons, however,

results in a nonzero percentage of simulated correlations with magnitudes larger than 0.45, when fixing the number of students at 15. Conversely, estimating value added with any number of students between 15 and 35 does not yield such correlations when fixing the number of lessons at 1. This difference in the effect on correlations caused by increasing students versus lessons is reflective of the differences in reliability of value-added and observational measures.

The above findings raise an important policy issue related to requirements for different types of teacher quality measures. It is not uncommon to see districts and states require a minimum numbers of student tests to inform a value-added estimate for it to be used in an evaluation. But while this is commonplace for value added, it is not uncommon for districts to rely on observation ratings that are based on a single classroom observation. Our findings, which are grounded in empirically-based parameters, suggest that the reliability benefits of adding an additional observation are far higher than the reliability benefits of marginal increases in the number of students that inform value-added measures.

We illustrate the magnitude of each parameter's effect on average correlations between value added and observations by estimating the following model using OLS regression:

$$(3)\ \rho_{\mathrm{VAM}',\mathrm{OBS}'} = \beta_0 + \beta_1(\rho_{\mathrm{VA,TQ}}) + \beta_2(\rho_{\mathrm{OBS,TQ}}) + \beta_3(\mathrm{N}_{\mathrm{VA}'}) + \beta_4(\mathrm{N}_{\mathrm{OBS}'}) + \beta_5(\mathrm{N}_{\mathrm{VA}'})^2 + \beta_6(\mathrm{N}_{\mathrm{OBS}'})^2 + \varepsilon$$

The outcome of Equation (3), $\rho_{\mathrm{VAM}',\mathrm{OBS}'}$, represents the average correlations between value added and observations, estimated with error, across the 100 simulations, for each combination of parameters for measure validity and reliability. The coefficients from this regression can be seen below in **Table 2**.

[Table 2 about here]

Table 2 shows the pattern of the effect of reliability on correlations that we saw earlier: the main effect of increasing the number of lessons used to estimate observations with error is

larger than the main effect of increasing the number of students used to estimate value added. However, the effect of increasing the number of lessons also demonstrates greater diminishing returns. The coefficients in Table 2 also allow us to compare the relative effects of measure validity and reliability on correlations, despite the two sets of parameters being measured on different scales. For example, the small effect of number of lessons or number of students on correlations, coupled with the improbability of collecting a substantial number of data points to estimate teacher effectiveness measures in practice, suggests that measure (in)validity may largely drive the weak relationships seen in research. On the other hand, we can calculate that even if both value added and observations were perfect proxies of the same dimension of teacher quality (i.e., $\rho_{\mathrm{VA,TQ}} = 1$ & $\rho_{\mathrm{OBS,TQ}} = 1$), the relationship between the two measures would be moderate if measure reliability was low (i.e., an average correlation of 0.406 if value added were estimated using 15 students, and observation scores were estimated using one lesson). Furthermore, increases in the validity of student test performance or observed lesson performance as measures of teacher quality largely cannot be realized in practice without changing the measures themselves; increasing measure reliability through the estimation of scores using more data points, however, is a lever that researchers or practitioners can utilize. These overarching patterns depicted in our results stress both the importance of measure validity and reliability when attempting to explain weak correlations in extant research.

In addition to examining the overall patterns of effects on correlations caused by changes to measure validity and measure reliability, we also explore the specific scenarios that likely lead to weak relationships. The first explanation for weak correlations seen in research between value added and observations is that both measure the same dimension of teacher quality, but estimates are unreliable. **Figure 3** below shows how average correlations between the two scores, observed

with error, change at different levels of measure reliability when both value added and observations serve as valid proxies for teacher quality (i.e., $\rho_{\mathrm{VA,TQ}} = 0.80$ & $\rho_{\mathrm{OBS,TQ}} = 0.80$).

[Figure 3 about here]

This figure depicts that, when both value added and observations correlate relatively strongly with the same underlying dimension of teacher quality, we largely observe a moderate relationship, falling within the range of existing findings (i.e., correlations between 0.30 and 0.45). When estimating value-added or observation scores with error at higher levels of reliability, however, we begin to witness correlations with magnitudes greater than most seen in prior research. This indicates that though in most cases we cannot reject the possibility that both value added and observations serve as good proxies for the same dimension of teacher quality, a weak relationships observed in practice or research between relatively more reliable measures likely signals that one or both measures are less valid.

The second and third explanations for weak correlations seen in research is that one or both reliably measure different constructs, related or unrelated to different dimensions of teacher quality. **Figure 4** below shows how average correlations between scores, estimated using the likely upper bound for reliability of measures in most teacher evaluation systems (i.e., scores estimated from 35 students and six lessons), change at different levels of measure validity.

[Figure 4 about here]

Results displayed in this figure also indicate that we cannot reject the second and third explanations for weak observed relationships. These scenarios suggest that the correlations between value added and observations should be weak or moderate when one or both measures serve as invalid proxies for the same dimension of teacher quality—the correlations depicted in the top-left, bottom-left, and bottom-right quadrants of the figure. In these quadrants, value

added and observations largely correlate with one another at magnitudes less than 0.15, within the range of prior findings.

5. Conclusions

Teachers' contributions to student test performance and classroom observations of teacher practice have become integral parts of updated teacher evaluation systems. Researchers and practitioners typically consider both value added and observations to measure the same underlying construct: teacher quality. Yet research has largely shown through correlations that the two measures only weakly relate to one another. This finding may be judged to be problematic for theoretical reasons, as effective instruction should lead to greater student learning, and student test performance should reflect student learning. Moreover, the weak relationship may also lead to perception issues that inhibit the feasibility of certain targeted job actions aimed at improving teacher practice; it could undermine the notion that one or both measures are accurate representations of true teacher quality.

Three different scenarios can, on their own or in conjunction, result in these weak correlations between measures. In the first scenario, both measures serve as good proxies for the same dimension of teacher quality, but in practice are estimated unreliably due to sampling error, attenuating correlations. In the second scenario, teacher quality is multidimensional, and even when value-added and observational measures are estimated reliably, the two measures capture different types of teacher quality. The third scenario is closely related to the second, except that one or both measures serve as proxies for constructs completely unrelated to teacher quality.

Results from our simulations did not allow us to rule out any of the scenarios for the weak correlations seen in prior research. For example, correlations between relatively more reliable scores were generally lower than 0.45 when one or both measures were correlated to the

same underlying dimension of teacher quality at less than 0.50. Yet correlations between value-added and observational measures fell within the range seen in many prior studies, such as the MET project (Kane & Staiger, 2012), even when both measures, estimated from a "typical" number of data points were highly related to the same underlying dimension of teacher quality. Investigating the particular scenarios again corroborated the importance of measure validity and highlighted the possibility of multidimensionality of measures, but also stressed the likelihood that both value added and observations actually do measure the same dimension of teacher quality, contrary to some recent propositions.

We acknowledge that there are limitations to our findings because we use simulated data in our analyses. Specifically, the conclusions we draw depend on the specific set of parameters we have chosen to model. In order to address this issue, we tested a large combination of parameters ($N = 45{,}486$) which encompassed numerous permutations of the factors that prior research has demonstrated to influence correlations between value-added and observations. Furthermore, we have chosen values for these factors that largely reflect what typical teacher evaluation systems might face (i.e., number of students used to estimate value added ranging from 15 to 35, or the number of lessons used to estimate observation scores ranging from one to six) and to reflect prototypical values in research. We believe that our simulations account for most possible conditions in which correlations between value added and observations can emerge in research or in practice, though future investigations may opt to select parameters more appropriate for their purposes. For example, middle school teachers evaluated using value-added measures may teach a larger numbers of students than the "upper-elementary school" teachers we simulate in our data.

Finally, we wish to highlight that the framework that we have used to illustrate the scenarios for why value-added and observational measures of teacher quality may be weakly related can also be applied to other areas where the relationship between two different measures is of interest. Future work in the field of education, in particular, might thus use simulated data to explore the likelihood of different scenarios that explain observed relationships between other teacher quality measures, such as student perception surveys, experience, knowledge, or preparation.

References

Amrein-Beardsley, A. (2014). VAMs and observations: Consistencies, correlations, and contortions. Retrieved from http://vamboozled.com/vams-and-observations-consistencies-correlations-and-contortions/

Bacher-Hicks, A., Chin, M., Kane, T. J., & Staiger, D. O. (2015). *Validating components of teacher effectiveness: A random assignment study of value-added, observation, and survey scores*. Manuscript in preparation. Harvard Graduate School of Education, Cambridge, MA.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., ... & Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. EPI Briefing Paper# 278. *Economic Policy Institute*.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

Brophy, J., & Good, T. L. (1986). Teacher Behavior and Student Achievement. *Handbook of Research on Teaching*, 328-375.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2014). Trends in classroom observation scores. *Educational and Psychological Measurement*, 0013164414539163.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*(5), 757-783.

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *The American Economic Review*, *104*(9), 2633-2679.

Cohen, J., & Goldhaber, D. (in press). What we know about teacher VAMs compared to other measures of evaluating teachers. In J.A. Grissom & P. Youngs (Eds.), *Making the most of multiple measures: The impacts and challenges of implementing rigorous teacher evaluation systems*. New York: Teachers College Press.

Corcoran, S., & Goldhaber, D. (2013). Value added and its uses: Where you stand depends on where you sit. *Education*, *8*(3), 418-434.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 8-15.

Doherty, K. M., & Jacobs, S. (2013). State of the states 2013 connect the dots: Using evaluations of teacher effectiveness to inform policy and practice. *National Center on Teacher Quality*.

Fox, L. (in press). Playing to Teachers' Strengths: Using multiple measures of teacher effectiveness to improve teacher assignments. *Education Finance and Policy*.

Gershenson, S. (in press). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., Whitehurst, G. J., & Croft, M. (2011). *Passing muster: Evaluating teacher evaluation systems*. Washington, DC: The Brookings Institution.

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics*, *7*(3), 199-208.

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the "Rothstein Falsification Test": Does it really show teacher value-added models are biased?. *Journal of Research on Educational Effectiveness*, *8*(1), 8-34.

Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects.*Economics of Education Review*, *36*, 216-228.

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

Goldhaber, D., & Theobald, R. (2013). Do different value-added models tell us the same things? *Carnegie Knowledge Network Briefs*. Stanford, CA.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, 0013189X14544542.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445-470.

Guarino, C. M., Reckase, M. D., Stacy, B. W., & Wooldridge, J. M. (2015). Evaluating specification tests in the context of value-added estimation. *Journal of Research on Educational Effectiveness*, *8*(1), 35-59.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, 280-288.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, *100*(1), 84–117.

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

Harris, D. N. (2012). How do value-added indicators compare to other measures of teacher effectiveness? *Carnegie Knowledge Network What we know series: Value-added methods and application.* Stanford, CA. Retrieved from http://www.carnegieknowledgenetwork.org/wp-content/uploads/2012/10/CKN_2012-10_Harris.pdf

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

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

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794-831.

Ho, A. D. & Kane, T. J. (2013). The reliability of classroom observations by school personnel. *Bill and Melinda Gates Foundation.*

Jackson, C. K. (2012). *Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina* (No. w18624). National Bureau of Economic Research.

Jennings, J. L., & DiPrete, T. A. (2010). Teacher effects on cultural capital development in elementary school. *Sociology of Education*, *83*, 135-159.

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

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains. *Policy and practice brief prepared for the Bill and Melinda Gates Foundation*.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587-613.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47-67.

Lynch, K., Chin, M., & Blazar, D. (2015). *Relationships between observations of elementary teacher mathematics instruction and student achievement: Exploring variability across districts.* Manuscript submitted for publication.

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2013). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement*, 0013164413515882.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67-101.

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

McClellan, C., Donoghue, J., & Park, Y. S. (2013). Commonality and uniqueness in teaching practice observation. *Clowder Consulting.*

Mehta, N. (2015). *Targeting the wrong teachers? Linking measurement with theory to evaluate teacher incentive schemes. Manuscript in preparation*. University of Western Ontario, London, Ontario.

Minnici, A. (2014). The Mind Shift in Teacher Evaluation: Where We Stand--and Where We Need to Go. *American Educator*, *38*(1), 22-26.

Murnane, R. (1981). Interpreting the evidence on school effectiveness. *The Teachers College Record*, *83*(1), 19-35.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects?. *Educational evaluation and policy analysis*, *26*(3), 237-257.

Papay, J. P. (2011). Different Tests, Different Answers The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal*, *48*(1), 163-193.

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123-141.

Polikoff, M. S. (2014). Does the test matter? Evaluating teachers when tests differ in their sensitivity to instruction. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teahcing Project* (pp. 278-302). San Francisco, CA: Jossey-Bass.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 247-252.

Rothstein, J. (2008). *Teacher quality in educational production: Tracking, decay, and student achievement* (No. w14442). National Bureau of Economic Research.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537-571.

Rothstein, J. (2014). Revisiting the impacts of teachers. *Unpublished working paper. http://eml. berkeley. edu/~ jrothst/workingpapers/rothstein_cfr. pdf*.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in education*, 8(3), 299-311.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247-256.

Sass, T. R. (2008). The stability of value-added measures of teacher quality and implications for teacher compensation policy. Brief 4. *National Center for Analysis of Longitudinal Data in Education Research*.

Schacter, J., & Thum, Y. M. (2004). Paying for high-and low-quality teaching. *Economics of Education Review*, *23*(4), 411-430.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). Sage Publications.

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

Figures and Tables

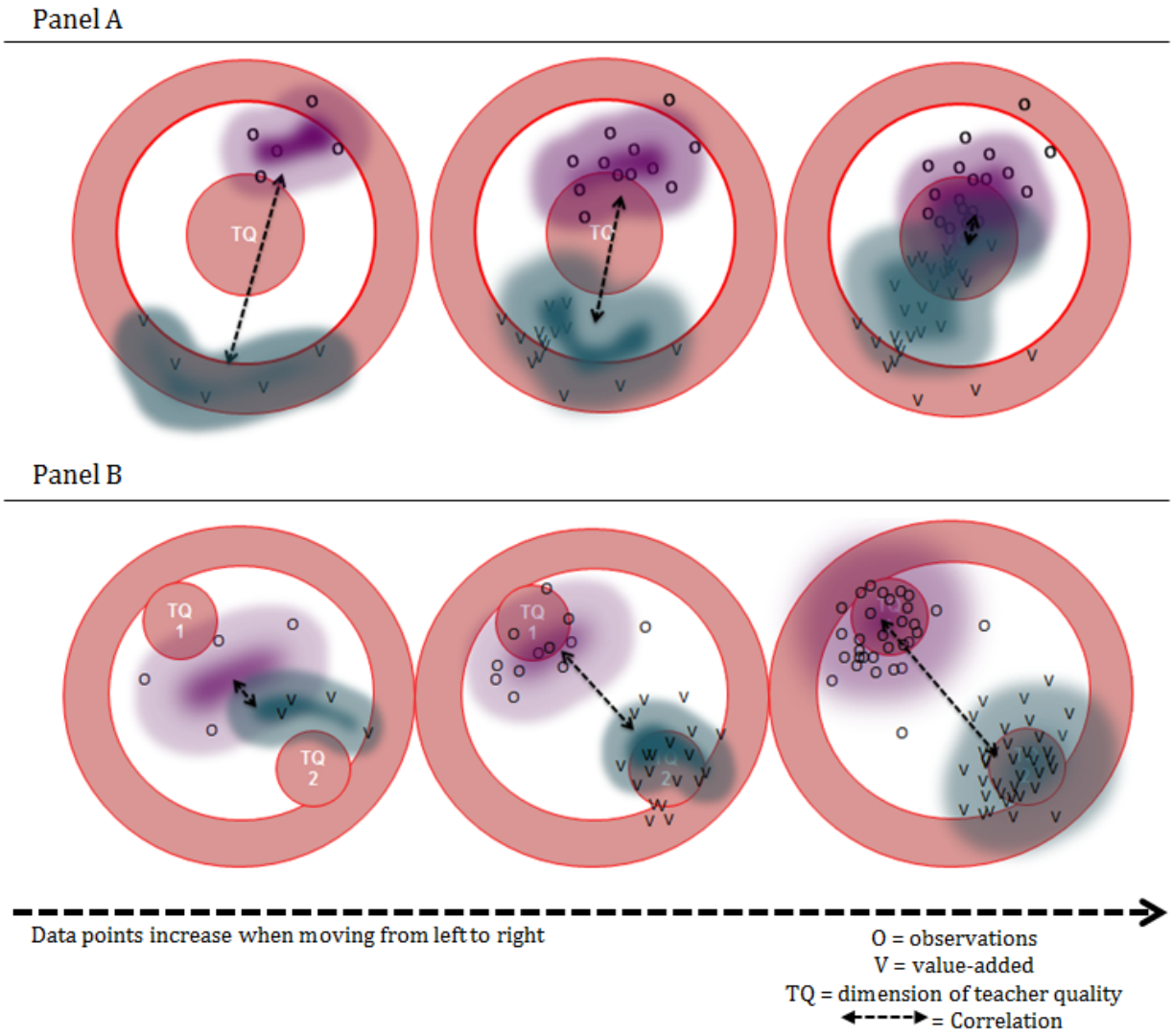

Figure 1. Possible scenarios of relationships between value-added and observation scores.

Table 1. Description of Varying Parameters of Different Simulations

Category	Parameter	# Combos	Description	Symbol
Multidimensionality / Validity	Correlation between TQ and error-free value added	19	Correlations range from 0.05 to 0.95, at intervals of 0.05	$\rho_{\mathrm{VA,TQ}}$
	Correlation between TQ and error-free observational quality	19	Correlations range from 0.10 to 0.90, at intervals of 0.05	$\rho_{\mathrm{OBS,TQ}}$
Reliability	Number of students used to estimate value added	21	Number of students range from 15 to 35	$\mathrm{N}_{\mathrm{VA}'}$
	Number of lessons used to estimate observational quality	6	Number of lessons range from 1 to 6	$\mathrm{N}_{\mathrm{OBS}'}$

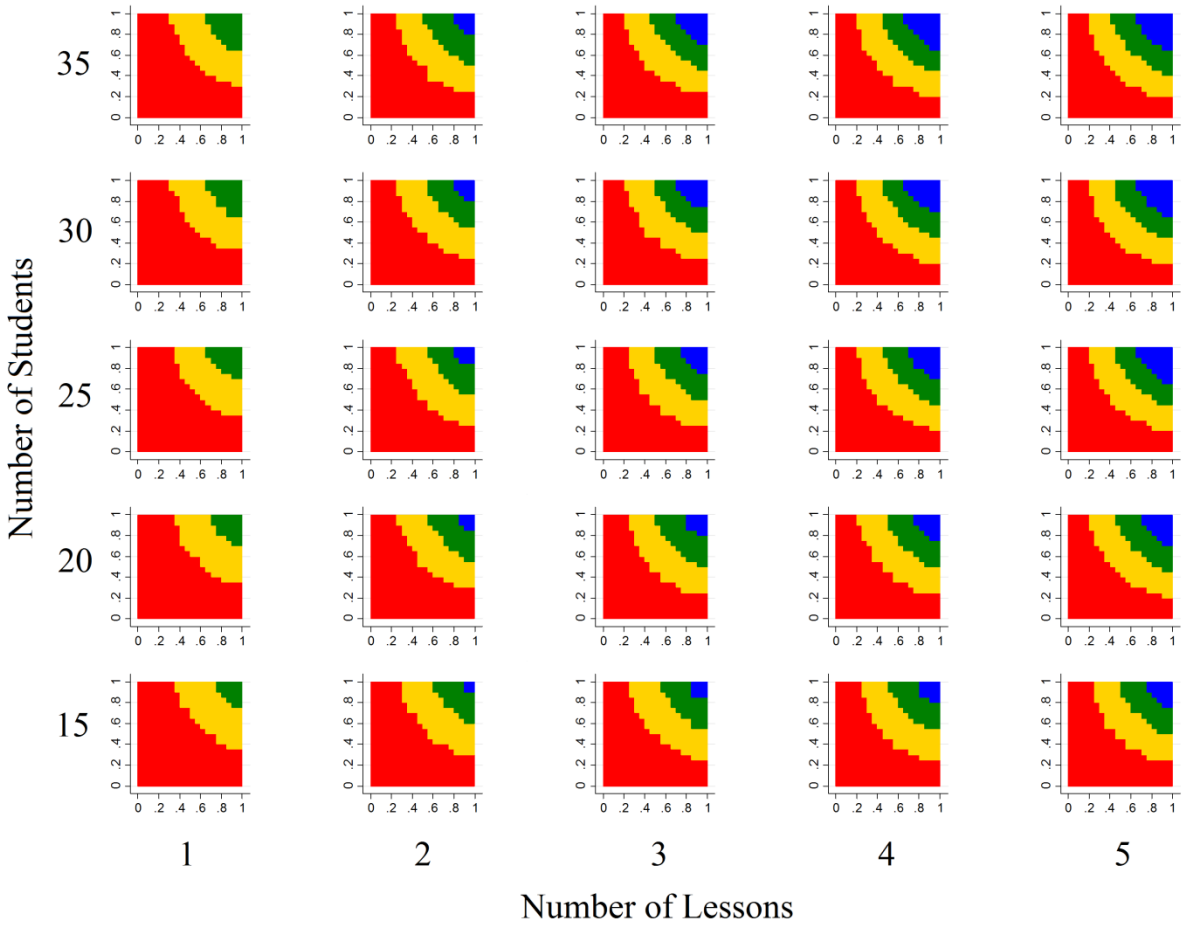

Figure 2. Magnitude of average correlations between VA and OBS, with error, depicted. For each subgraph, the y-axis is the correlation between TQ and error-free VA, and the x-axis is the correlation between TQ and error-free OBS. Red regions represent correlations between VA and OBS less than 0.15. Yellow regions represent correlations between VA and OBS between 0.15 and 0.30. Green regions represent correlations between VA and OBS between 0.30 and 0.45. Blue regions represent correlations between VA and OBS greater than 0.45.

Table 2. Regression coefficients predicting average correlations

	$\rho_{\mathrm{VAM}',\mathrm{OBS}'}$
Measure Validity	
$\rho_{\mathrm{VA,TQ}}$	0.300***
	(0.001)
$\rho_{\mathrm{OBS,TQ}}$	0.301***
	(0.001)
Measure Reliability	
$\mathrm{N}_{\mathrm{VA}'}$	0.003***
	(0.000)
$(\mathrm{N}_{\mathrm{VA}'})^2$	-3.31×10^5***
	(0.000)
$\mathrm{N}_{\mathrm{OBS}'}$	0.030***
	(0.001)
$(\mathrm{N}_{\mathrm{OBS}'})^2$	-0.003***
	(0.000)
β_0	-0.260***
	(0.004)
Number of Unique Parameters	45,486

Note: Standard errors reported in parentheses. ****p<0.001.*

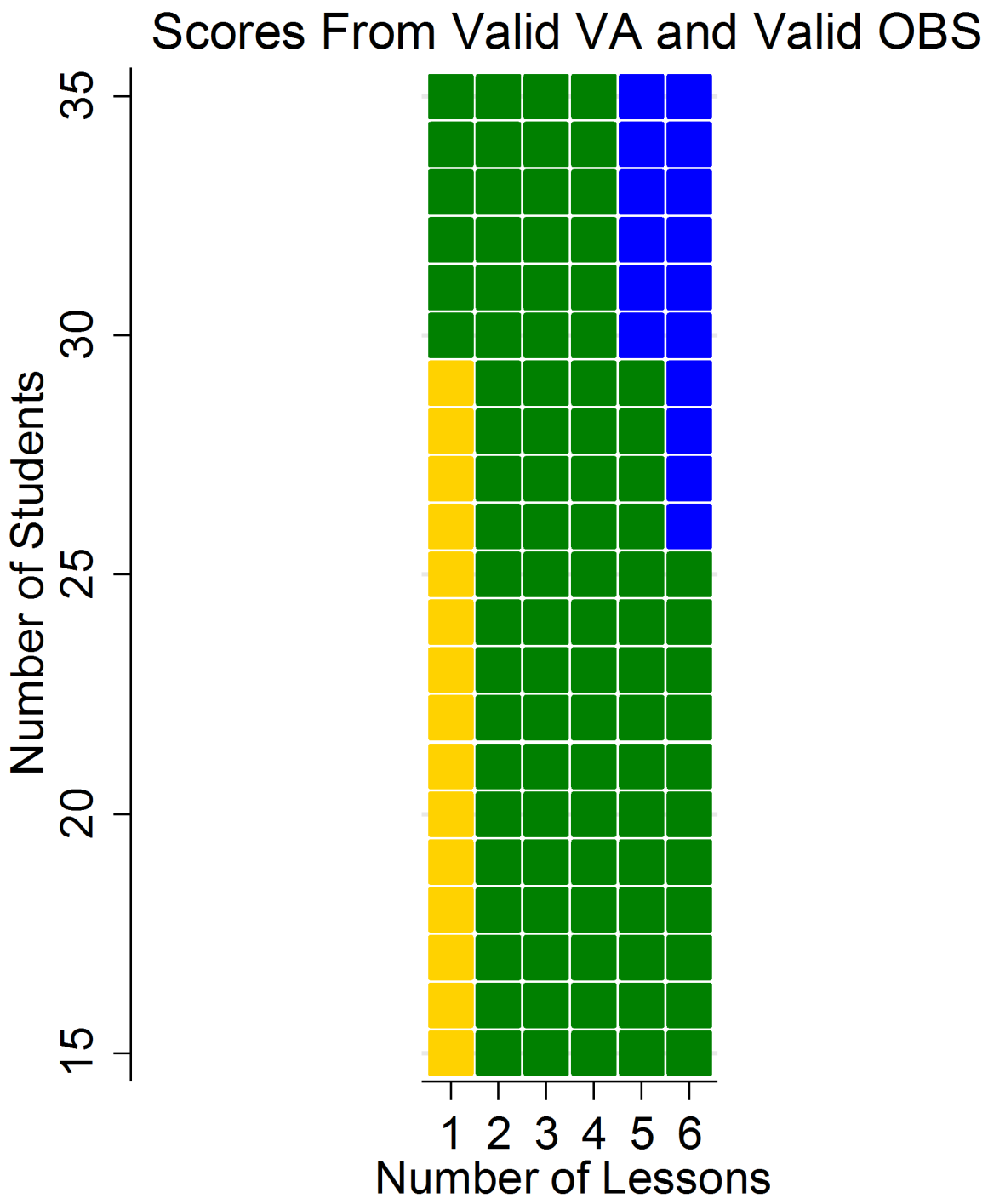

Figure 3. Magnitude of average correlations between VA and OBS, with error, depicted. Underlying correlation between error-free VA and TQ is 0.80. Underlying correlation between error-free VA and OBS is 0.80. Red regions represent correlations between VA and OBS less than 0.15. Yellow regions represent correlations between VA and OBS between 0.15 and 0.30. Green regions represent correlations between VA and OBS between 0.30 and 0.45. Blue regions represent correlations between VA and OBS greater than 0.45.

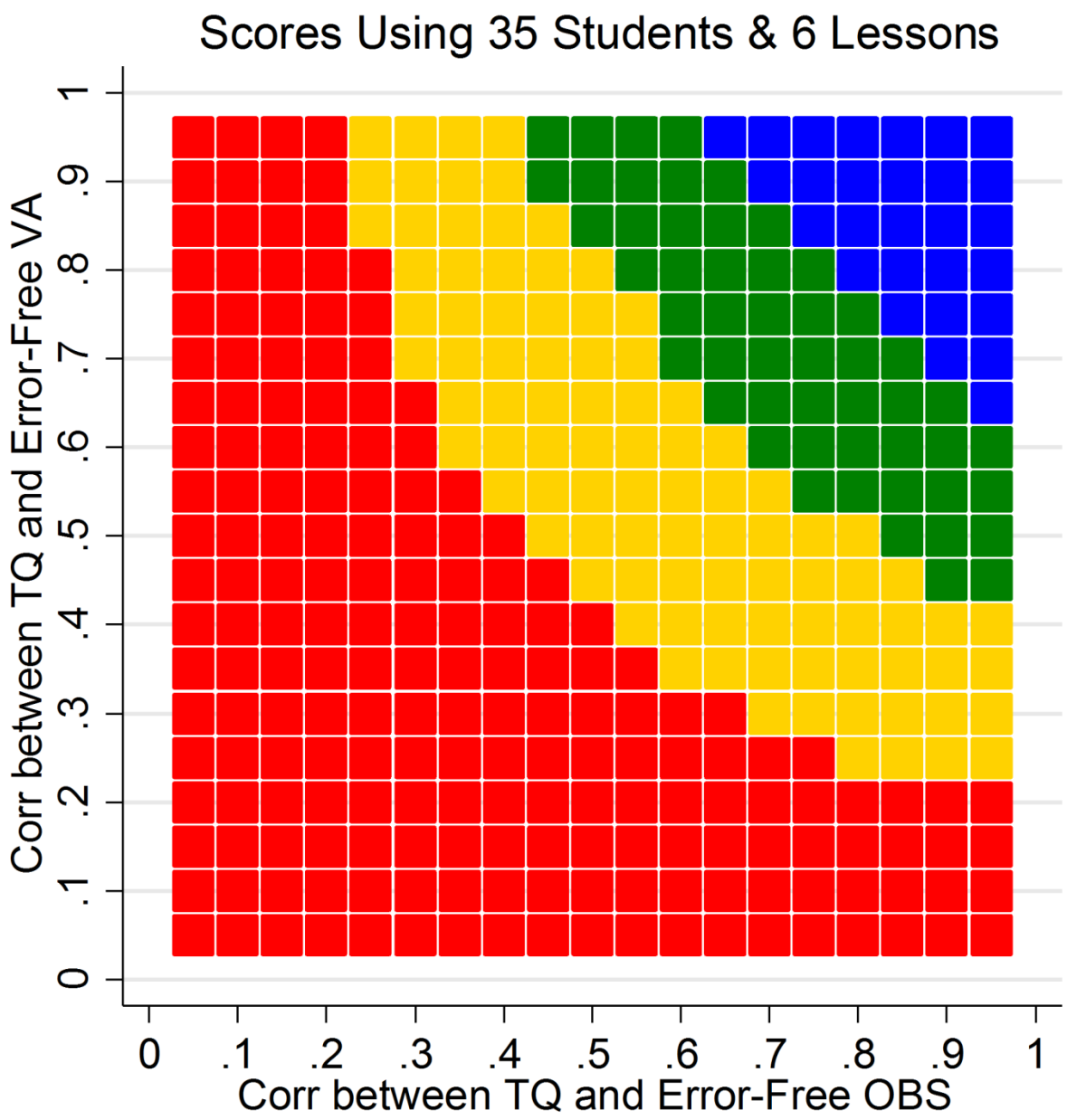

Figure 4. Magnitude of average correlations between VA and OBS with error depicted. Red regions represent correlations between VA and OBS less than 0.15. Yellow regions represent correlations between VA and OBS between 0.15 and 0.30. Green regions represent correlations between VA and OBS between 0.30 and 0.45. Blue regions represent correlations between VA and OBS greater than 0.45.