Making Value-Added Inferences from Large-Scale Assessments

Derek C. Briggs

University of Colorado at Boulder

October 15, 2011

Pre-print from Simon, M., Ercikan, K., & Rousseau, M (Eds.) (2012) *Improving Large-Scale Assessment in Education: Theory, Issues and Practice*. London: Routledge

Summary

Value-added inferences about student learning will often be an unavoidable consequence when student achievement on a large-scale assessment is summarized with respect to teachers or schools. A distinguishing feature of value-added models is the availability of longitudinal data containing, at a minimum, current and prior year information about student test scores in a given subject domain. This chapter reviews and explores conceptual and pragmatic factors that are likely to influence the validity and reliability of value-added inferences about teacher or school quality.

Introduction

As longitudinal data from large-scale assessments of academic achievement has become more readily available in the United States, educational policies at the state and federal levels have increasingly required that such data be used as a basis for holding schools and teachers accountable for student learning. If the test scores of students are to be used, at least in part, to evaluate teachers and/or schools, then a key question to be addressed is "how"? To compare them with respect to the average test score levels of their students would surely be unfair because we know that student test performance is strongly associated with variables that are beyond the control of teachers and schools—prior learning, the social capital of a child's family, etc. The premise of value-added modeling is to level the playing field in such comparisons by only holding teachers and schools accountable for the student learning that it would be possible for them to influence. For example, imagine the availability of test scores in mathematics for some population of students across two grades, 4 and 5. For each student, performance on the math test in grade 5 is predicted as a function of other variables known to be associated with academic achievement. These variables would include, at a minimum, performance on the grade 4 math test, but might also include test performance in earlier grades and other test subjects, and information such as whether or not students were eligible for free and reduced lunch services, were English language learners, etc. Finally, the grade 5 test score that has actually been obtained for any given student can be subtracted from the score that was predicted. These values can be aggregated at the classroom or school level. If the resulting number is positive, it could be considered evidence that a teacher or

school has, on average, produced a positive effect on student learning. If negative, a negative effect could be presumed. This approach, known as value-added analysis, is increasingly being promoted as a fair and objective way to make judgments about teacher (and/or school) effectiveness (Gordon, Kane, & Staiger, 2006; Harris, 2009).

The purpose of the present chapter is to provide the reader with a survey of key issues that need to be grappled with before one engages in the process of making value-added inferences from large-scale assessments. The use of value-added modeling to evaluate the quality of teachers and schools in the United States is controversial. Many have expressed alarm at what they perceive as an unbridled enthusiasm for the approach given the technical limitations that are inherent to it (Baker et al, 2010; Braun, Chudowsky, & Koenig, 2010). Others, while acknowledging the concerns, argue that even with its deficiencies, value-added analyses represent a marked improvement relative to preexisting approaches to educational accountability (Glazerman et al., 2010). Though I have my own opinions on these matters (Briggs, 2008; Briggs & Domingue, 2011), and these will probably become evident to the reader, I attempt to maintain an objective perspective throughout. There are five sections that follow. I begin by defining the distinguishing features of value-added models, and go into some detail to present the functional form of two widely used approaches. (For those readers uninterested in the esoteric details of these models, these subsections can be skipped without losing the narrative thread of the chapter.) The second section addresses the question of whether, in a statistical sense, value-added models can be used to make direct causal inferences about the effects of teachers and schools on student achievement. There has been some heated academic debate on this topic, and I summarize key points of contention. In the third

section, I present the perspective that value-added estimates are best interpreted as descriptive indicators rather than casual effects. In doing so I draw an analogy between the approach taken by the father of epidemiologiy, John Snow, to attribute the cause of a cholera outbreak to contaminated drinking water, and the approach that would be required to attribute a value-added indicator to a teacher or school. The fourth section addresses the issue of the stability of value-added estimates. To the extent that these estimates are clouded by measurement error, the value-added signal may be overwhelmed by noise. I describe a number of approaches commonly taken to adjust for measurement error. In the fifth section, I briefly take note of five implicit assumptions typically made about large-scale assessments that serve as the key inputs before a value-added output can be computed: alignment with standards, curriculum and instruction; interval scale properties; vertical links across grades; unidimensionality; and constant measurement error. I conclude the chapter with a summary of the major points that have been raised.

Value-Added Models

For other more comprehensive reviews of value-added modeling the interested reader should consult the following monographs and published articles: (Braun et al., 2010; Hanushek & Rivkin, 2010; Harris, 2009; McCaffrey et al., 2003; McCaffrey, Han, & Lockwood, 2009; OECD, 2008). In this section I begin by focusing attention on the distinguishing features that separate value-added models from other psychometric or statistical models that are sometimes prevalent in the assessment and evaluation of student learning (see for example, Zumbo, Wu & Liao (this volume)). In the National

Research Council report *Getting Value out of Value-Added*, Braun et al. (2010) define value-added models (VAMs) as "a variety of sophisticated statistical techniques that use one or more years of prior student test scores, as well as other data, to adjust for preexisting differences among students when calculating contributions to student test performance." (Braun et al. 2010, 1) According to Harris (2009), "the term is used to describe analyses using longitudinal student-level test score data to study the educational input-output relationship, including especially the effects of individual teachers (and schools) on student achievement." From these definitions, two key features of VAMs are implicit. First, all VAMs use, as inputs, longitudinal data for two or more years of student test performance. Second, VAMs are motivated by a desire to isolate the impact of specific teachers or schools from other factors that contribute to a student's test performance. It follows from this that the output from a VAM is a numeric quantity that can be used to facilitate causal inferences about teachers or schools. In what follows I will refer to these numeric quantities as value-added *estimates*, or alternatively, as value-added *indicators*.

The Production Function Approach

Let Y_{ig} represent an end of year test score on a standardized assessment for student (*i*) in grade (*g*), in a classroom with teacher (*t*) and school (*s*). The VAM is specified as

$$Y_{igts} = \alpha_g + \beta X_{ig} + \gamma Z_{ig}^{(ts)} + \sum_t \theta_t D_{ig} + \varepsilon_{igts}. \quad (1)$$

The covariates in this model are captured by X_i which represents test scores from prior grades[1] , and $Z_{ig}^{(ts)}$, which could represent any number of student, classroom or school-specific variables thought to be associated with both student achievement and classroom assignment. In this particular specification of the model, the key parameter of interest is the fixed effect q_t, which represents the effect of the current year's teacher on student achievement (i.e., the model above includes a dummy variable indicator for each teacher in the dataset). The model could also be written such that fixed effects for schools were the key paramerers of interest.

The validity of the model hinges upon the assumed relationship between the unobserved error term e_{igts} and q_t. If, conditional upon *X* and *Z*, e_{igts} and q_t are independent, in theory it is possible to obtain an unbiased estimate of q_t. In other words, if one can control for the variables that govern the selection process whereby higher or lower achieving students land in certain kinds of classrooms, then one can approximate the result that would be obtained if students and teachers had been randomly assigned to one another from the outset. This is controversial proposition, and much of the debate over the use of VAM for teacher accountability has focused on (a) the nature and number of covariates that need to be included in X and Z, and (b) evaluating the extent to which adding more variables or student cohorts serves to reduce bias in $\hat{q}_t$.

The production function model has a long history in the economics literature (Hanushek & Rivkin, 2010; Todd & Wolpin, 2003), and helps to explain why this has

[1] Depending on the current year grade of the student, the number of available prior test scores in the same subject could range anywhere from 1 (if current year grade of student i is 4), to 8 (if current year grade of student *i* is 12).

been the preferred specification approach among economists who have contributed the VAM literature[2].

The EVAAS Approach

The Educational Value-Added Assessment System (EVAAS; Sanders, Saxton & Horn, 1997) has the longest history as a VAM used for the purpose of educational accountability. While a detailed presentation is outside the scope of this chapter, a key point of differentiation between it and the approach presented above can be seen by writing out the equation for a single test subject in parallel to equation 1

$$Y_{ig} = a_g + \sum_{g^* \leq g} q_{g^*} + e_{ig}. \quad (2)$$

In contrast to the fixed effects specification from the production function approach, the EVAAS represents a multivariate longitudinal mixed effects model. As such, teacher effects for a given grade are cast as random variables with a multivariate normal distribution such that $q_{g^*} \sim N(\mathbf{0}, t)$. Only the main diagonal of the covariance matrix is estimated (i.e., teacher effects are assumed to be independent across grades). The student-level error term is also cast as a draw from a multivariate normal distribution with a mean of 0, but the covariance matrix is left unstructured. The EVAAS is often referred to as the "layered model" because a student's current grade achievement is expressed as a cumulative function of the current and previous year teachers to which a student have been exposed. For example, applying the model above to longitudinal data that span grades 3 through 5 results in the following system of equations:

[2] For example, this is the specification that underlies the VAM used by the University of Wisconsin's *Value-Added Research Center*, which has taken an active role marketing its services to urban school districts across the country (e.g., New York City, Milwaukee, Los Angeles).

$$Y_{i3} = a_3 + q_3 + e_{i3}$$
$$Y_{i4} = a_4 + q_3 + q_4 + e_{i4}$$
$$Y_{i5} = a_4 + q_3 + q_4 + q_5 + e_{i5}$$

In the model above no teacher effects can be computed for grade 3 because they are confounded with variability in student achievement backgrounds. In contrast, when certain assumptions hold it is possible to get an unconfounded effect for the grade 4 teacher. This can be seen by substituting the first equation into the second equation in the system such that $Y_{i4} - Y_{i3} = a_4 - a_3 + q_4 + e_{i4} - e_{i3}$. This shows that the sufficient statistic for estimates of teacher effects under the EVAAS are test score gains from one grade to the next. It is for this reason that the EVAAS (and other mixed effect modeling approaches related to it) has long been presumed to require test scores that had been vertically scaled (Ballou, Sanders, & Wright, 2004; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004).[3]

This simplified presentation may obscure two significant aspects of the EVAAS that contribute to its purported ability to validly and reliably "disentangle" the influence of teachers from other sources of that influence student achievement. In particular, the EVAAS

- makes use of panel data for up to five years of longitudinal cohorts per teacher simultaneously; and
- models multiple test subject outcomes jointly as a multivariate outcome.

Though it has been criticized because it does not control for additional covariates beyond a student's test score history (Kupermintz, 2003), teacher value-added estimated by the

[3] As it turns out, while it can matter how a test has been scaled, it will generally make little difference to value-added rankings of teachers or schools whether the tests have or have not been vertically linked 6/28/2013 2:09:00 PM.

EVAAS with and without student-level covariates have been shown to be strongly correlated (Ballou, Sanders & Wright, 2004). A more equivocal issue is whether or not one should control for classroom or school-level characteristics (McCaffrey et al, 2004). Controlling for classroom characteristics can lead to overadjustments of teacher effect estimates; controlling for school characteristics will restrict teacher comparisons to a within-school reference population. Finally, note that the EVAAS assumes that the effects of students' teachers in the past persist undiminished into the future. McCaffrey et al (2004) and Lockwood, McCaffrey, Mariano & Setodji (2007) have demonstrated empirically that this may not be a viable assumption in the context of teachers; Briggs & Weeks (2011) show that it is probably not viable in the context of schools.

How Well Do VAMs Estimate the Effects of Teachers or Schools on Student Achievement?

Historically most educational research has focused on interventions that are "manipulable" from a policy perspective. By manipulable it is meant that it would be relatively easy (though not necessarily cheap) to expose students to more or less of the intervention. Examples would include reductions in class size, the introduction of web-based learning technologies to a curriculum, and test-based grade retention. What has made the value-added approach simultaneously intriguing and controversial has been a shift in focus since the late 1990s to define teachers and schools themselves as the principle educational interventions of interest.

When a VAM is used to estimate the effect for a more traditionally manipulable educational intervention, the intended interpretation is relatively straightforward. Indeed, the average causal effect of a manipulable intervention has a natural "value-added" interpretation: it is the amount by which a student's test score outcome differs from what it would have been in the absence of the intervention (i.e., the counterfactual outcome). In contrast, Rubin, Stuart, & Zanutto (2004) and Raudenbush (2004) have pointed out that the value-added estimates associated with teachers and schools are very difficult to conceptualize in a similar, counterfactually meaningful way. This is largely because as an educational intervention, the amalgamated characteristics of the teachers and/or schools to which a student is assigned are difficult to change—i.e., to manipulate—over a finite period of time. The more technical objection to the causal interpretation of a value-added estimate is that such estimates are likely to be biased because students and teachers are sorted—or sort themselves—into classrooms and schools on the basis of many different variables that are plausibly related to how students perform on achievement tests. For some variables that are observable, such as a student's prior academic achievement, free or reduced lunch status, parent's education level, or English language proficiency, it may be possible to control for confounding statistically as part of a regression modeling approach. However, to the extent that much of this sorting occurs because of unobserved variables that might be known to a school's principal but not to a value-added modeler (e.g., student motivation, parental involvement, teacher rapport with students, etc.), there is usually good reason to suspect that estimates of a teacher or school effect will be biased.

In a sequence of papers Rothstein, (2009; 2010) demonstrated, using data from North Carolina, that teacher effects from three commonly specified VAMs were significantly biased due to student and teacher sorting. Rothstein (2010) accomplished this by implementing a simple yet ingenious "falsification" check. Logically, it would be impossible for a teacher in the future to have an effect on student achievement gains in the past. If this were found to be the case, a plausible explanation is that the model is capturing the impact of one or more omitted variables from the model that is correlated with teacher assignments. As a result, the effect of these omitted variables is being erroneously attributed to a future teacher. Therefore, the argument goes, the omission of these variables will also lead to the erroneous attribution of effects to current year teachers. By showing that commonly used VAM specifications (e.g., subsets of equation 1 above) failed his falsification check, Rothstein was able to call into question the premise that VAMs are capable of disentangling the effects of teachers from the effects of other factors on student achievement. Rothstein (2009) also found that under certain simulated scenarios in which students are assigned to teachers by principals on the basis of variables that would not be available to the value-added modeler, the magnitude of bias in estimates of teacher effectiveness could be substantial.

When Rothstein's falsification check has been applied to test the validity of similar VAM specifications with data from Los Angeles and San Diego Unified School Districts, others have found similar evidence of bias in the value-added estimates (Briggs & Domingue, 2011; Koedel & Betts, 2011). However, some researchers appear to have reached more optimistic conclusions about the validity of value-added inferences when using more complex VAM specifications to evaluate the effectiveness of teachers and

schools. For example, although Koedel & Betts (2011) corroborated Rothstein's principal finding, they also found that evidence of bias could be greatly mitigated by averaging teacher value-added estimates across multiple cohorts of students while also restricting the comparison to those teachers who have shared the same students over time. This argument hinged upon the notion that sorting is transitory from year to year—a teacher may get a favorable class one year but not necessarily the next. In the San Diego schools Koedel & Betts examined this seems to have been the case to some extent. Whether this assumption would hold in other contexts is less clear.

The most optimistic case for the validity of teachers effects estimated using value-added models comes from an randomized experiment that was conducted by Kane & Staiger (2008) usng a sample of 78 pairs of classrooms from Los Angeles schools. In their study, Kane & Staiger were able to estimate value-added effects for teachers before they were randomly assigned to classrooms in a subsequent year. They argued that if value-added can be used to accurately distinguish effective from ineffective teachers, then a significant proportion of the variability in classroom to classroom achievement levels at the end of the experiment should be attributable to differences in prior year estimates of teacher value-added. Their results were supportive of this hypothesis. By the end of the school year in which teachers had been randomly assigned, within school differences in classroom student achievement could be accurately predicted from prior differences in the estimates of value-added for the respective teachers in each classroom. On the basis of these results Kane & Staiger argued that "conditioning on prior year achievement appears to be sufficient to remove bias due to non-random assignment of

teachers to classroom (p. 3)" a finding that would seem to directly contradict the more pessimistic conclusions reached by Rothstein.

There are, however, several relevant criticisms of the Kane & Staiger study and the implication that even a very simple VAM could be used to support direct causal inferences about teacher effectiveness. First, because elementary schools self-selected to participate in the original experiment, it would be plausible that the principals at these schools agreed to participate because teachers are already assigned to classrooms in a manner that is very close to random before the experiment began. To the extent that this is true, then the weight placed upon a VAM to adjust statistically for biases due to the purposeful sorting of teachers and students is lessened considerably. Second, Rothstein (2010) has argued that because the experiment involved a relatively small number of classrooms and teachers, it did not have sufficient statistical power to detect significant discrepancies between the observed effects attributed to teachers, and that which was predicted by the VAM. Third, in no case did Kane & Staiger show evidence that a VAM specification which *only* conditioned on prior achievement was an accurate predictor of "true" differences in teacher effectiveness. Rather, all VAM specifications in the Kane & Staiger study controlled for both prior achievement and a number of other demographic characteristics at both the student and the classroom levels. It follows then, that even if one were to agree that Kane & Staiger had successfully isolated the effects of teachers from other possible factors, the VAMs used to accomplish this were, in fact, quite complex.

This issue seems particularly relevant when the results from a VAM are to be used to directly classify teachers into categories of effectiveness; in some cases

differences in the variables included in the model can lead to significant differences in teacher classifications. As a case in point, on August 14, 2010, just prior to the start of the 2010-11 academic school year, the *Los Angeles Times* published results from a value-added analysis of elementary schools and teachers in the Los Angeles Unified School District (Felch, Song & Smith, 2010). The data for this analysis came from multiple cohorts of students taking the reading and math portions of the California Standardized Test between 2003 and 2009. The VAM used to estimate teacher effects consisted of a regression model that controlled for five student-specific variables: prior year test performance, gender, English Language Learner status, enrollment in a school receiving Title 1 funding, and whether or not the student was enrolled in the district as of Kindergarten (Buddin, 2010). Based directly on the estimates from this model, teachers were placed into normative quintiles of "effectiveness", and these ratings were made publicly available at a dedicated web site. Yet in a re-analysis of the same data, Briggs & Domingue (2011) showed that the VAM being used by the *Los Angeles Times* not only failed the Rothstein falsification test, but that if additional control variables had been included in the model, 54% and 39% of grade 5 teachers would have changed effectiveness classifications in reading and math respectively.

A major driver of these differences in classifications was the decision by Briggs & Domingue to include average classroom achievement (intended as a proxy for peer influence) and additional prior test score information as control variables for each student in the data. Hence it could be argued that the biggest problem with the *Times's* decision to publish value-added ratings of teachers was that the VAM they had applied was not sufficiently complex to adequately support a valid causal inference. But even this

position can be problematic. For example, as was noted previously, the decision to include classroom or school-level control variables for achievement and/or poverty in a VAM can lead to overadjustment if, for example, more effective teachers tend to seek jobs in higher-achieving and affluent schools (McCaffrey et al., 2004, Ballou, Sanders & Wright, 2004, and OECD, 2008). And though it has been established that the inclusion of multiple years of prior test scores can reduce the bias in estimated teacher effects (Rothstein, 2010), such specifications would typically preclude the use of a VAM for teachers in the early years of elementary school.

In summary, the question of whether VAMs can be used to make direct causal inferences about teacher effectiveness is still a matter of ongoing debate. For more optimistic perspectives see Goldhaber & Hansen, 2010; Sanders, Saxton & Horn, 1997; Sanders & Horn, 1998; for more tempered perspectives see Ishii & Rivkin, 2009; McCaffrey et al., 2003; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Reardon & Raudenbush, 2009. While there are probably few researchers that would argue that VAMs can fully disentangle the influence of the teacher from all other factors that contribute to student achievement[4], there are many who are convinced that certain VAMs can at least validly distinguish between those teachers who are at the low and high extremes of a hypothetical effectiveness distribution. At the time this chapter was being written, a large-scale replication of the Kane & Staiger's value-added experiment in Los Angeles was being conducted in five urban school districts across the country (the

[4] William Sanders and his colleagues behind the EVAAS may be a notable exception. For example, SAS makes the following explicit claim in its marketing of its Educational Value Added Assessment System: "It is much more than teacher or classroom level analyses; it assesses the effectiveness of districts, schools and teachers, as well as provides individual student projections to future performance. SAS EVAAS for K-12 provides precise, reliable and unbiased results that other simplistic models found in the market today cannot provide." Retrieved August 25, 2011 from http://www.sas.com/govedu/edu/k12/evaas/index.html.

"Measures of Effective Teaching" project funded by the Gates Foundation). To the extent that the results of this replication corroborate the initial findings from Kane & Staiger's original study, this would lend support to the use of VAMs as a central element of high-stakes evaluations. However, while the results from this study will be eagerly anticipated, it seems unlikely that these results will settle the issue unequivocally.

Value-Added Inferences and Shoe Leather

Should the numeric outcome from a value-added analysis be interpreted as direct estimates of teacher or school effectiveness, or, alternatively, as descriptive indicators that lead only indirectly to inferences about effectiveness? This subtle distinction may explain why there continues to be confusion over the distinction between a growth model and a VAM. Like VAMs, growth models depend upon the availability of longitudinal data on student test performance. Unlike VAMs, many growth models (see Zumbo, Wu & Liu, this volume) are used first and foremost to support descriptive and exploratory analyses in which inferences about students, rather than classrooms or schools, are of primary interest. The key distinction between growth and value-added—inferential intent—is easy to blur because the moment that student-level growth statistics are summarized at the classroom or school levels, it will often be unavoidable that high or low values are attributed to teachers or schools in a causal manner. Yet while all growth models applied to students in educational settings may well encourage direct causal inferences about teacher effectiveness, in the absence of additional evidence, there may be no compelling theoretical reason to believe that these inferences are valid. At the

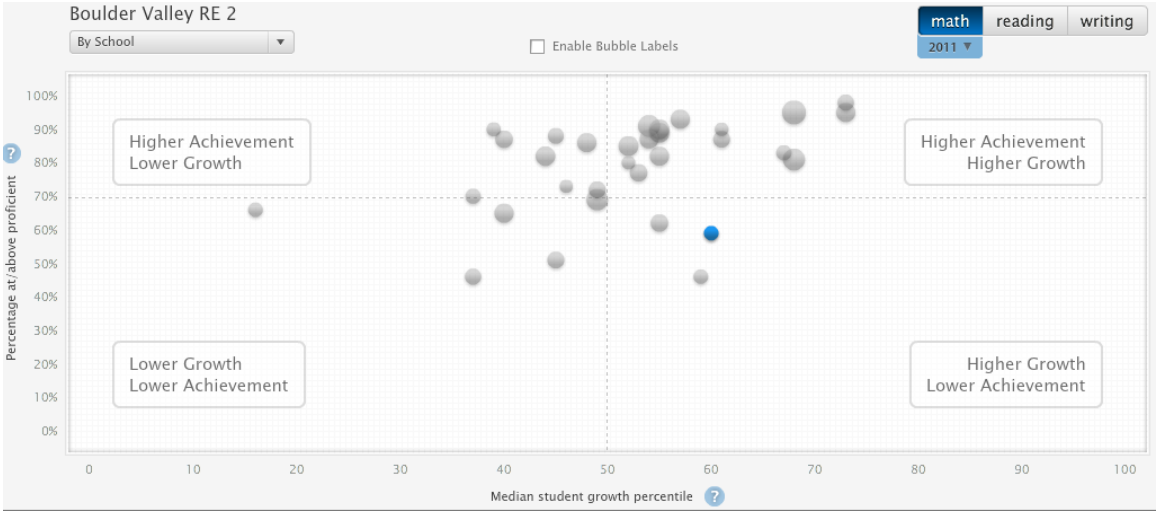
same time, the fact that value-added models are intended to support direct inferences about teacher and school effects does not necessarily make them any more valid than growth models when both are used for the same purpose.

Because of the difficulties inherent in claiming that a statistical model can be used to estimate the effect of a teacher or school on student achievement, some have argued that the outputs from a value-added analysis are best interpreted as descriptive indicators (Briggs, 2008; Hill, Kapitula, & Umland, 2010). The interpretation of a classroom or school-level statistic as a descriptive indicator rather than a causal effect shifts the technical conversation from a consideration of *internal validity* to a consideration *construct validity*. That is, if a numeric quantity associated with a teacher is to be interpreted as a causal effect, the fundamental validity issue from a statistical point of view is whether we can obtain parameter estimates that are unbiased and precise. In contrast, if that same number is to be interpreted as a descriptive indicator, the fundamental validity issue is the extent to which empirical evidence can be provided that collectively supports the intended interpretation and use of the indicator (c.f., Kane, 2006). The latter task is just as challenging as the former, and is decidedly messier and much less proscriptive as a process. Nor does it mean that issues of causal inference can or should be avoided. If a descriptive measure is a significant source of evidence being used to reward or sanction teachers, the implied inference that, for example, higher quality teaching is associated with higher values of the descriptive indicator would need to be defended empirically. Nonetheless, the labeling of teacher or school growth statistics as fundamentally descriptive communicates the message that more evidence

must be gathered before any high-stakes decisions on the basis of these indicators are warranted.

To illustrate this, we now look more closely at the educational accountability context in the state of Colorado. (In this example we will focus on schools rather than teachers as the principal units of analysis.) The student growth percentile model (Betebenner, 2009) and its accompanying graphical interface (www.Schoolview.org) have become the backbone of Colorado's approach to educational accountability. In short, the student growth percentile model computes for each student a conditional test score percentile. This conditional score percentile is found by, in essence, comparing the test score performance of a student in a given grade to all students in the state with the same test score history in all prior grades. A student that scores at the 50^{th} percentile of this conditional distribution is one that is inferred to have shown "growth" that represents "one year of learning." An estimate of classroom or school-level growth can then be computed by taking the median over all student growth percentiles for students with test scores in at least two adjacent grades. Note that in this approach the concept of growth is an inference—we infer that if a student has performance that is higher than expected relative to similar students, the reason for this is that they have learned more (i.e., shown more growth) than these similar students. This stands in contrast to the growth models described by Zumbo, Wu & Liu (this volume), all of which are focused on making growth inferences in terms of changes in absolute magnitudes.

Figure 1. Plots of School-Level Achievement and Growth, Boulder Valley School District, 2011



The bubbleplot in Figure 1 shows how the student growth percentile model is likely to promote value-added inferences about school quality. Each "bubble" in the plot represents a unique elementary school in the Boulder Valley School District. The size of each bubble is proportional to the number of students attending a given school. The vertical axis represents the percentage of students in a given school that have been classified as proficient or advanced on Colorado's large-scale assessment in mathematics. In contrast, the horizontal axis represents a school's median student growth percentile. In this example, the vertical line at 50 (the median) represents the demarcation between schools that are performing better or worse than would be expected given the prior test performance of their students. The horizontal line at 70 represents the threshold set by

Colorado for a school to be considered "high achieving." Figure 1 make it possible to distinguish among four "types" of schools:

1. Quadrant I: Higher Achievement, Higher Growth
2. Quadrant II: Higher Achievement, Lower Growth
3. Quadrant III: Lower Achievement, Lower Growth
4. Quadrant IV: Lower Achievement, Higher Growth

The frames of reference for the descriptors "higher" and "lower" are the vertical and horizontal thresholds that have been established by the Colorado Department of Education. The further a school departs from these thresholds, the clearer the designation of a school within each quadrant. The school in the plot with the bubble shaded in blue is an example of a school that greatly benefits from taking a two-dimensional perspective on academic success. From the perspective of achievement status, the students at this school are not on pace with what is expected of them academically. Yet relative to their prior achievement, they appear to be learning a significant amount of mathematics.

The student growth percentile model is a great example of why the terms growth model and value-added model are sometimes used interchangeably: because student growth percentiles can be easily aggregated to the classroom and school-levels, they are often given a de facto attribution as value-added. In the example above it would be difficult to sidestep the inference that "effective" and "ineffective" schools are those with median growth percentiles above and below 50, respectively. Are such inference problematic? This will depend on the process through which such inferences lead to direct sanctions or rewards as part of an overarching system of educational accountability. At one extreme, we could imagine a scenario in which value-added

estimates are combined across available subject domains (i.e. math and reading) and then used to classify schools into categories of effectiveness. Schools found to be ineffective would be sanctioned, schools found to be effective rewarded. At another extreme, value-added estimates would be combined with scores from holistic rubrics used to assess (for example) school climate through direct observation. A composite index might be formed, and this, along with purely qualitative information (e.g., a principal's narrative account, parent interviews, student work products) would be used in the attempt to root out the causes of any positive or negative trends being observed. (These same two extremes could just as easily be envisioned when teachers rather than schools are being evaluated. See Hill et al, 2010 for a thorough illustration.) In the first scenario, one would probably be hard-pressed to defend the use of something like the student growth percentile model to make an immediate and direct causal attribution. In contrast, for the second scenario, a value-added estimate is no longer being directly attributed as the effect of a school on student achievement. Instead, it is one of multiple sources of information being used in what will ultimately require a subjective judgment about school quality by personnel at the district or state levels. In this second scenario, the question of whether or not the value-added estimate is biased in a statistical sense may no longer be the salient issue so long as its use as part of a larger pool of evidence that leads to desirable and defensible outcomes.

The Role of Shoe Leather

The eminent statistician David Freedman was notoriously pessimistic about the notion that statistical modeling, in and of itself, could lead to direct causal attributions (Freedman, 1987, 1991, 2004, 2006).

> I do not think that regression can carry much of the burden of a causal argument. Nor do regression equations, by themselves, give much help in controlling for confounding variables…I see many illustrations of technique but few real examples with validation of the modeling assumptions (Freedman, 1991, 292).

Freedman would often argue that it was not statistical modeling in isolation, but the use of statistical reasoning supported primarily by old-fashioned detective work that led to important breakthroughs in causal inference. The most famous example Freedman was known for relaying was that of the father of epidemiology, John Snow, credited for establishing the causal link between polluted drinking water and death by cholera.

During a devastating outbreak of cholera in London's Soho neighborhood in the summer of 1854, Snow was able to trace the root cause of the outbreak to the contaminated water from the nearby Broad Street pump in Soho's Golden Square. At the time of the outbreak, Snow was in the process of tabulating the results from a natural experiment in which a subset London residents had been exposed—for all intents and purposes at random—to the drinking water from two companies, one of which (Lambeth) had moved their site for water acquisition to an unpolluted section of the Thames River, and another of which (Southark & Vauxhall) had not. Sometime after the cholera outbreak that devastated the Soho neighborhood, Snow was able to establish through the quantitative results of his "grand experiment" that the drinking of polluted drinking water was the principal cause of cholera. However, at the time of the 1854 outbreak, Snow did

not yet have this evidence at his disposal. He was able to build a strong correlational case nonetheless by creating a street map of the area surrounding the Broad St pump and showing that there was a strong association between mortality rates and distance from the pump. The force of his argument convinced the Board of Governors of St. James Parish to remove the handle from the pump, and this almost surely averted a subsequent outbreak (Johnson, 2006).

Snow's success in making a convincing causal attribution in this instance did not come from using a statistical model that could disentangle the effect of drinking well-water from other possible causes (e.g., miasma, unsanitary living conditions, etc.). Rather, it was possible because he had spent years developing a hypothesis that cholera was a waterborne agent contracted not by breathing polluted air (the predominant hypothesis at the time) but by eating or drinking contaminated food or water. Over the course of a decade, Snow had developed increasingly convincing tests of his hypothesis while simultaneously ruling out alternate hypotheses. By the time of the cholera outbreak at Golden Square, Snow already had a good idea where to look for the root cause. Snow gathered evidence to support his theory by knocking on doors and interviewing the remaining residents (even as most of them were fleeing the scene). What Snow learned in the process was not gleaned from the ingenious specification of a regression model, but from wearing down his shoe leather as he made the rounds in the streets of London.

Of course, one can only go so far in drawing the analogy between the process of attributing death by cholera to the drinking of contaminated water and the process of attributing variability in student achievement to the quality of schools or teachers. In the value-added context the outcome of interest, student learning, is a latent construct that is

decidedly more difficult to measure than death! And while it was ultimately possible to validate Snow's theory at the microscopic level[5], it will never be possible to establish definitively that a given school or teacher was responsible for what a collection of students have or have not learned.

But certain parallels are relevant. What can be established through a value-added analysis is suggestive evidence that certain schools and teachers are associated with greater gains in student achievement than others. This is not that much different from Snow showing that proximity to the Broad St pump was strongly associated with mortality rates. To go from association to a convincing argument for causation, Snow had to track down and explain away all the instances that worked against his theory (e.g., residents who had died without drinking from the Broad St pump, residents who had drunk from the Broad St pump and not died). One would expect the same approach to unfold when value-added is used to evaluate schools or teachers. If a teacher's students have, on average, performed worse than was expected, this can be cast as a hypothesis as to the quality of the teaching students have received. To ignore this evidence altogether would be irresponsible. On the other hand, there may be competing hypotheses from other sources of evidence that lead to different inferences. Some of this evidence may be entirely qualitative (e.g., interviews with parents, peer observations of teaching practice), or it may come from quantitative scrutiny (e.g., the statistical model used to generate the value-added estimate failed to account for a critical variable). The key point is that almost irrespective of the particular statistical approach taken to generate a value-added

[5] The Italian scientist Filippo Pacini is credited with the discovery of *Vibrio cholera* in 1854, but his discovery appears to have been ignored by the scientific community at the time. It was the German physician Robert Koch who later won more widespread acceptance for the etiology of cholera after his (independent) discovery of the bacterium in 1884.

indicator, the validity of high-stakes decisions about teachers and schools will depend upon the way that these indicators are understood by stakeholders, and the extent to which they facilitate the kind of detective work that is needed to make a causal attribution.

The Stability of Value-Added Indicators

Measurement error can pose problems for stability of value-added indicators in two different ways at two different levels. First, to the extent that the student-level regression equations at the foundation of a value-added analysis includes prior year test scores as "control" variables, measurement error in these observed scores will lead to an attenuation of *all* regression coefficients in the model (Fuller, 1987; Buonocarsi, 2010). Second, regardless of the quality of instruction to which they are exposed, it may be the case that some cohorts of students are simply "better" or "worse" than others[6]. Given this, when teachers or schools are the units of analysis to which inferences are being made, it has been argued that some portion of the observed variability in estimates of value-added can be explained by chance (Kane & Staiger, 2002; McCaffrey, Sass, Lockwood, & Mihaly, 2009). The key distinction here is that measurement error at the student level is assumed to have a functional relationship with the number of test items that students have been administered; at the teacher or school level, measurement error is assumed to have a functional relationship with the number of students. The analogy here is essentially that students are to schools what items are to students. Taken together, both

[6] This is sometimes described as "sampling error" rather than measurement error, but the concept is the same.

of these sources of measurement error could explain the phenomenon in which value-added estimates appear to "bounce" up and down in a volatile manner from year to year—even if the true value were actually constant over time.

Measurement Error at the Classroom or School Levels

The year to year correlation of teacher value-added has been found to be weak to moderate, ranging from about 0.2 to 0.6 (Goldhaber & Hansen, 2008; McCaffrey et al., 2009). Kane & Staiger (2002; 2008) have argued that such intertemporal correlations can be interpreted as an estimate of reliability, in which case any intertemporal correlation less than 0.5 would imply that more than half of the variability in value-added can be explained by chance unrelated to characteristics of teacher or school quality that persist over time. After conducting a simulation study, Schochet & Chiang (2010) conclude that 35% of teachers are likely to be misclassified as either effective or ineffective when classifications are based on a single year of data.

Three adjustments are typically made, sometimes in tandem, to account for the instability of value-added indicators. One adjustment is to increase the number of years of data over which value-added is being computed. Schochet & Chiang (2010) find that going from one to three years of data reduces the error rate for teacher effectiveness classifications in their simulation from 35 to 25%. Using empirical data from Florida, McCaffrey et al. (2009) find that going from one to three years of data increases, on average, the reliability of value-added estimates for elementary school and middle school teachers from 0.45 to 0.55 and 0.56 to 0.66 respectively. Two related adjustments are to use these estimates of reliability to "shrink" value-added estimates back to the grand

mean (i.e., the average value-added of all teachers in the system), and/or to compute a confidence interval around each value-added estimate.

When the reliability of value-added is low, it will typically be a mistake to attempt to classify teachers or schools into more than three categories (e.g., significantly below average, average, significantly above average). In such instances if teachers are instead classified into quintiles of the value-added distribution (five equally spaced categories instead of three unequally spaced categories), misclassification rates are likely to increase dramatically (Aaronson, Barrow, & Sander, 2007; Ballou, 2005; Briggs & Domingue, 2011; Koedel & Betts, 2007).

Measurement Error at the Student Level

Considerably less research has been done to evaluate the impact of student-level measurement error on value-added analyses. The problem only appears to be relevant to regression models in which prior year test scores are included as independent variables. Such cases lead to the classic "errors in variables" problem that is well-understood in the econometrics literature. Though there are many possible adjustments that could be used as a correction to this problem (c.f., Fuller, 1987; Buonacorsi, 2010), the adjustments that have been applied in the literature to date (c.f., Buddin & Zamarro, 2009; Rothstein, 2009) have been based on the assumption of constant measurement error across the test score distribution, an assumption that is clearly unrealistic given the way that large-scale assessments are designed (see next section). While it is clear that the failure to adjust for the error in variables problem can have a significant impact on value-added inferences,

the practical impact of imposing linear instead of nonlinear adjustments is unclear. This is likely to be an active area for research studies in the coming years.

Implicit Assumptions about Large-Scale Assessments

Test Validity

Value-added analyses tend to be agnostic about the quality of the underlying assessments use to generate test score outcomes. But if the underlying assessments are judged to be invalid because of either construct underrepresentation or construct-irrelevant variance, then it makes little difference whether a value-added estimate can be used to reliably isolate the direct effect of teachers or schools on student achievement. Much of this will hinge upon the alignment between the standards and expectations of what students should know and be able to do from grade to grade, the curriculum and instruction to which students are exposed, and the questions they are asked when they take a summative assessment at the end of the school year. Along these lines, teachers and schools can only be expected to have discernible effects on student achievement when test items are sensitive to relatively short-term instruction. Yet not all tests are designed with this intent. As just one example, the items on the SAT are designed to measure quantitative and verbal critical reasoning that is developed gradually over the course of many years of schooling (Messick, 1980). In contrast, the sorts of tests used within systems of educational accountability are presumed to be sensitive to high-quality instruction. This constitutes a fundamental assumption that has seldom been formally investigated.

Test Scaling

With the exception of the student growth percentile model, the models used to conduct value-added analyses implicitly assume that test score outcomes exist on an interval scale. This means that score differences should have the same meaning in an absolute sense whether the initial score came from, for example, the low or high ends of the scale. Ballou (2009) argues that this assumption is both implausible and impossible to evaluate empirically. He suggests that value-added analyses would be on surer footing if they were to use statistical methods that only required data with ordinal properties. Briggs (2010) has argued that good empirical methods do exist for psychometricians to evaluate whether test scores have interval properties, but that such methods and their importance are not well-understood or appreciated by most psychometricians. For more on this issue, see Michell (1997, 1999, 2000, 2008).

Vertical Linking

Value-added analyses can be conducted whether or not test scores have been placed on a vertically linked score scale. For models in which prior year test scores are used as control variables this is readily apparent. But even for models that use repeated measures as the outcome of interest, the presence or absence of a linking step after scores have been scaled within a grade has been shown to have a negligible impact on normative value-added rankings (Briggs & Weeks, 2009b; Briggs & Betebenner, 2009). It can be shown that the presence or absence of a vertical link between the test scores from adjacent grades will only have an impact of value-added ranking when there is a

substantial change (e.g., increase or decrease of more than 20%) to the variability of test scores from grade to grade. This is not to argue that vertical scales are not valuable for other reasons—if carefully designed they offer great promise for more meaningful criterion-referenced interpretations about growth magnitudes (Stenner, Burdick, Sanford & Burdick, 2006; Stenner & Stone, 2010). But for the purpose of making value-added inferences, a vertically linked scale is not necessary.

Unidimensionality

Test scores from large-scale assessments are ultimately composites of the different components of knowledge, skill and ability that comprise a measurement construct. In some cases it may be reasonable to treat this composite as a single coherent dimension; in other cases this may lead to a loss of important information about what students know and can do. Lockwood et al. (2007) examined four years of longitudinal data for a cohort of 3,387 students in grades 5 through 8 attending public schools in the state of Pennsylvania from 1999 to 2002. Of interest was the sensitivity of teacher effect estimates to the complexity of the VAM being specified. The authors chose four different VAMs in order of the complexity of their modeling assumptions. They also chose five different sets of control variables to include in the VAMs: none, demographics, base year test score, demographics plus base year test score, and teacher-level variables. Finally, they considered one novel factor seldom explored in prior VAM sensitivity analyses: the dimensionality of the outcome measure. Students in the available sample had been tested with the Stanford 9 assessment across grades 5 through 8. Upon examining the items contained in the Stanford 9, Lockwood et al. disaggregated the test into two different

subscores as a function of items that emphasized problem solving (40% of the test) and items that emphasized procedures (60% of the test). Having established three factors for their sensitivity analysis (type of VAM, choice of covariates, choice of test dimension), the authors estimated teacher effects for each three-way factor combination and asked the question: Which factor has the greatest impact on inferences about a given teacher's effect on student achievement? What they found was that, by far, the choice of test dimension had the biggest impact on teacher effect estimates. Regardless of the choice of VAM or covariates, estimates of teacher effects tended to be strongly correlated (0.8 or higher). On the other hand, the correlations of teacher effects estimates by outcome were never greater than 0.4, regardless of the underling VAM or choice of covariates. This suggests that violations of the assumption of unidimensionality can have a significant impact on value-added inferences.

Measurement Error

The most commonly applied adjustments for measurement error in VAMs assume (1) that observed test scores are a linear function of two independent components, the "true" value of interest and chance error, and (2) that the variance of the chance error is a constant. This assumption does not mesh well with the way that modern assessments are designed and maintained using item response theory (IRT). Specifically, a major distinction between IRT and classical test theory is that measurement error can be expressed as a function of a student's location on the score scale. Because students and test items can be placed on the same scale in IRT, it is easy enough to see that measurement error is always lowest at locations of the scale where there are the most

items. A usual consequence of this is that measurement error curves for large-scale assessments will tend to follow a U shape in which the magnitude of error is minimized in the middle of the scale, where the bulk of the test items and respondents tend to be located. At the extremes of the scale, where there are both fewer respondents and fewer items, measurement error will tend to be larger. Put in more general terms, if a large-scale assessment features relatively few items that would be considered very easy or very difficult for test-takers, then it will be hard to measure the lowest and highest performing students with the same precision as those students who are closer to the center of the score scale (e.g., within two standard deviations)[7]. All this implies that traditional adjustments for measurement error in value-added contexts are likely to be wrong when the adjustment relies solely upon a transformation of a summary statistic for score reliability such as Cronbach's alpha. Just how far off the adjustment will be is less clear, and will depend upon two factors, the shape of the measurement error curve (i.e., the test information function), and the proportion of students located at the tails of the distribution where measurement error is the largest.

Summary

Value-added inferences are unavoidable when large-scale assessments are used for summative purposes. In this chapter, different perspectives have been presented with respect to the way that estimates of value-added can be used to evaluate the quality of instruction that students receive from year to year. One perspective focuses primary

[7] An alternative way of casting this issue is with respect to floor or ceiling effects. Koedel & Betts (2010) have argued that ceiling effects will tend to bias the observed variability in teacher effects from a VAM downward.

attention on how we can isolate the effect of teachers and schools through statistical modeling and adjustment. When taking this perspective, the key issue that must be grappled with is whether statistical adjustments adequately account for the non-random nature by which teachers and students find themselves in different schools and classrooms. A second perspective—not necessarily incompatible with the first—is that the direct attribution of value-added to teachers or schools will always by equivocal, no matter how complex the underlying model. Given this, value-added should be viewed primarily as a descriptive indicator, and inferences based on these indicators can only be validated by referencing other sources of information, much of which may be qualitative in nature. In short, the observation that a teacher has a low or high value-added "score" in a given year raises the hypothesis that the teacher was ineffective or effective. While it would be irresponsible to ignore this evidence, it would also be irresponsible to avoid the subsequent detective work that would be necessary to reject or fail to reject the initial hypothesis. When taking this perspective, a case can be made that even less sophisticated growth models would be sufficient as the initial basis for a value-added inference, so long as the model takes the prior achievement of students into account and is transparent to stakeholders.

The stability and persistence of value-added estimates is important to consider irrespective of the underlying model used to compute them. Measurement error at the student level and at the classroom or school level can pose problems, and may explain why value-added estimates appear to bounce up and down from year to year. While statistical adjustments for measurement error at both levels are possible, they are not straightforward, and when scrutinized they may well turn out to be inadequate. Given

our limited understanding of how chance processes interact create noise that clouds the value-added signal, it will usually be advisable to be conservative when initially classifying teachers or schools as "effective" or "ineffective" on the basis of their value-added estimates.

Finally, while large-scale assessments are not—and should not—be designed according to the criterion of supporting value-added inferences, if the validity of the underlying tests is suspect, so to will be the validity of the value-added inferences that derive from them. Much is often implicitly assumed about the psychometric properties of test scores before they become the fundamental ingredients of a VAM: alignment with standards, curriculum and instruction; interval scale properties; vertical links across grades; unidimensionality; and constant measurement error. Violating one or more of these assumptions will not necessarily render value-added inferences meaningless (in fact, as was pointed out, the assumption of a vertical link is not needed), but it is a mistake to ignore assumptions being made about the underlying test scores altogether.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95-135.

Baker, E., Barton, P. E., Haertel, E., & F, H. (2010). Problems with the use of student test scores to evaluate teachers. Economic Policy Institute.

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In R. Lissitz (Ed) *Value added models in education: Theory and applications*, 272–303.

Ballou, D. (2009). Test scaling and value-added measurement. *Education finance and policy*, *4*(4), 351–383. MIT Press.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37-65.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42-51.

Braun, H., Chudowsky, N., & Koenig, J. (2010). *Getting Value Out of Value-Added*. Washington, DC: National Academies Press.

Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level value-added. *Journal of Educational and Behavioral Statistics*.

Briggs, D. C. & Domingue, B. D. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District Teachers by the Los Angeles Times. National Education Policy Center. http://nepc.colorado.edu/publication/due-diligence.

Briggs, D. C. (2010). The problem with vertical scales. Paper presented at the 2010 Annual Meeting of the American Educational Research Association, Denver, CO, May 3, 2010.

Briggs, D. C. & Betebenner, D. (2009) Is growth in student achievement scale dependent? Paper presented at the invited symposium "Measuring and Evaluating Changes in Student Achievement: A Conversation about Technical and Conceptual Issues" at the annual meeting of the National Council for Measurement in Education, San Diego, CA, April 14, 2009.

Briggs, D. C., & Weeks, J.P. (2009a). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*(4), 3–14.

Briggs, D. C., & Weeks, Jonathan P. (2009b). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy*, *4*(4), 384-414.

Briggs, D. C. (2008) The goals and uses of value-added models. Paper prepared for a workshop held by the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation and Educational Accountability sponsored by the National Research Council and the National Academy of Education, Washington DC: November 13-14, 2008.

Buddin, R. (2010). How Effective Are Los Angeles Elementary Teachers and Schools? Unpublished Manuscript. Retrieved December 15, 2010 from http://documents.latimes.com/buddin-white-paper-20100908/

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics*, *66*(2), 103-115.

Buonaccorsi, J. P. (2010). Measurement Error: Models, Methods, and Applications. New York: Chapman and Hall/CRC.

Felch, J., Song, J., & Smith, D. (2010, August 14). Who's teaching L.A.'s kids? *Los Angeles Times*. Retrieved February 2, 2011, from http://www.latimes.com/news/local/la-me-teachers-value-20100815,0,2695044.story

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational and Behavioral Statistics*, *12*(2), 101-128.

Freedman, D. A. (1991). Statistical models and shoe leather. *Sociological methodology*, *21*, 291-313.

Freedman, D. A. (2004). Graphical models for causation, and the identification problem. *Evaluation Review*, *28*(4), 267-293.

Freedman, D. A. (2006). Statistical models for causation: what inferential leverage do they provide? *Evaluation review*, *30*(6), 691-713.

Fuller, W. A. (1987). *Measurement Error Models*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Glazerman, S., Loeb, S., Goldhaber, D. D., Raudenbush, S., Whitehurst, G. J., & Policy, B. I. B. C. E. (2010). *Evaluating teachers: The important role of value-added*. *New York*. Brown Center on Education Policy at Brookings.

Goldhaber, D., & Hansen, M. (2008). Is it just a bad class? Assessing the stability of measured teacher performance. CRPE Working Paper 2008_5. Seattle, WA: Center on Reinventing Public Education.

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

Gordon, R., Kane, T., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job*. The Brookings Institution.

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

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy*, *4*(4), 319–350.

Hill, H. C., Kapitula, L., & Umland, K. (2010). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794-831.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy*, *4*(4), 520-536.

Johnson, S. (2006). *The Ghost Map: The Story of London's Most Terrifying Epidemic--and How It Changed Science, Cities, and the Modern World*. Riverhead Books.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), Educational Measurement (4th ed.). (pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, *16*(4), 91-114.

Kane, T., & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *NBER working paper*. Retrieved from http://www.nber.org/papers/w14607.

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

Koedel, C., & Betts, J. R. (2007). Re-Examining the Role of Teacher Quality In the Educational Production Function. Working Paper. Columbia, MO: University of Missouri

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, *25*(3), 287.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007a). Bayesian Methods for Scalable Multivariate Value-Added Assessment. *Journal of Educational and Behavioral Statistics*, *32*(2), 125-150.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007b) The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*. *44*(1), 47-68.

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.

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

McCaffrey, D.F. (2003). *Evaluating value-added models for teacher accountability* (Vol. 158). RAND Research Report prepared for the Carnegie Corporation.

McCaffrey, D.F., Han, B., & Lockwood, J. (2009). Turning student test scores into teacher compensation systems.

Messick, S. (1980). The Effectiveness of coaching for the SAT: review and reanalysis of research from the Fifties to the FTC. Princeton, Educational Testing Service: 135.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–384.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge University Press Cambridge, England.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*(5), 639.

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research & Perspective*, *6*(1), 7–24.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*, 23.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*(1), 121-129.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492-519.

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

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

Rubin, D. B., Stuart, E. a, & Zanutto, E. L. (2004). A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103-116.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system, a quantitative, outcomes-based approach to educational measurement. In Jason Millman (Ed.). Grading teachers, grading schools, Is student achievement a valid evaluation measure? (pp. 137-162). Thousand Oaks, CA: Corwin Press.

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.

Schochet, P. Z., & Chiang, H. S. (2010). Error rates in measuring teacher and school performance based on student test score gains. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Stenner, J., Burdick, H., Sanford, E., & Burdick, D. (2006). How accurate are lexile text measures? *Journal of Applied Measurement*, 7(3), 307-322.

Stenner, J., & Stone, M. (2010). Generally objective measurement of human temperature and reading ability: some corollaries. *Journal of Applied Measurement*, 11(3), 244-252.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*, 3-33.

Wright, S. P. (2010). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education. SAS White Paper.