Are value-added models good enough for teacher evaluations? Assessing commonly used models with simulated and actual data

GARY T. HENRY

RODERICK A. ROSE

University of North Carolina at Chapel Hill

ABSTRACT

Teachers' evaluations in many states include information about their students test score gains. In this paper, we describe the assumptions that are required for teacher value-added (TVA) estimates to be treated as unbiased causal effects. We compare commonly used TVA models on policy-relevant criteria using simulated data in which the assumptions of unconfounded assignment of students and teachers and no peer effects are violated and with actual data. The three-level hierarchical linear performs best when either assumption is violated. For year-to-year consistency, the dynamic ordinary least squares model performs best. A common policy goal – identifying the lowest performing quintile of teachers—can be done with reasonable accuracy but between 3.2 and 9.3 percent of all teachers are misclassified.

A virtue of value-added models is that they provide a means for researchers who seek to isolate the contribution of an entity such as an educational program, school, or teacher to changes in students' test scores during a finite time period. That is, rather than the "status-based" models that focus on cumulative achievement as measured by a single test score at a specific time, value-added models estimate the contribution of an educational program, schools, or teachers to students' test score changes over a set period of time, often a year of schooling (Tekwe 2004). Value-added models have been popularized in studies of the "education production function" (Todd & Wolpin 2003) that seek to isolate the impacts of specific "educational inputs" such as teacher quality (Clotfelter, Ladd & Vigdor 2007, 2010; Boyd et al. 2006,

2009; Rockoff 2004; Kane, Rockoff & Staiger 2008, author), class size (Angrist & Lavy 1999; Rivkin, Hanushek & Kain 2005; Boozer and Rouse 2001; Krueger 1999; Hanushek 1999), or expenditures (Ferguson & Ladd 1996; Hanushek 1994, 1999; Hedges, Laine & Greenwald, 1994) to changes in student test scores.

As of 2013, 35 states required measures of student achievement or growth be used as a significant factor in teachers' evaluations (NCTQ 2013). Most states have adopted teacher value-added (TVA) models or student percentile growth models to estimate the effects of individual teachers on their students' test scores.[1] Advocates of applying teacher value-added (TVA) models to the evaluation of individual teachers (Sanders, Saxton & Horn, 1997; Gordon, Kane, & Staiger 2006; Harris, 2009) and those opposed to their use (Amrein-Beardsley & Collins, 2012; Baker, et al. 2010; Hill 2009; Amrein-Beardsley 2008) often raise both substantive and methodological issues in their arguments. The former say TVA estimates are fairer, more objective, and "evidence-based" in the assessment of teacher effectiveness than alternatives that could be substitutes for TVA scores, while the latter group raises questions about the validity, complexity, transparency, fairness, accuracy of TVA scores, and inconsistencies in scores that are found for the same teachers across time.

In the present study we address three research questions: 1) Are some TVA models better than others at estimating the effects of individual teachers when assumptions for estimating causal effects are violated? 2) Do some TVA models more accurately identify the lowest (and highest) performing teachers than others when the assumptions are violated? 3) Do TVA models differ in the year-to-year consistency of individual teacher effectiveness estimates? These questions focus on a parsimonious set of criteria for evaluating TVA models that can aid our thinking about the evaluative purposes for which TVA scores might productively be used.

The first question focuses our attention on the main function of TVA models: to produce accurate estimates of the effects of individual teachers on their students' test scores when assumptions for regarding the effects as causal are routinely violated. If the models are not robust to violations of the assumptions needed to consider the estimates as causal, which we review later in the paper, the teacher scores will not be useful measures of teachers' effectiveness for evaluation purposes. We test this in two ways: (1) assessing the correlation with the "true" effect; (2) examining the total error of the estimates including bias and variance (mean square error). The second question focuses on one of the primary purposes of TVA scores: to identify teachers who deserve rewards or sanctions. If TVA models do not correctly identify the least or most effective teachers, then actions such as the denial of tenure, dismissal, or merit pay may be directed at the wrong teachers and miss the intended teachers. Finally, addressing the final question will help us obtain purchase on the use of TVA scores in teacher performance incentives. If incentives are based on teachers' scores but the scores are subject to extreme fluctuations from one year to the next, then teachers may be less likely to believe that TVA scores accurately reflect their performance, which could undermine their utility as a performance incentive for teachers (Stipek, 2013).

[1] While student percentile growth models actually estimate teachers' rank on the distribution of achievement growth of their students, we include them in the discussion of teacher value-added models since they estimate actual versus expected growth and are used for the evaluation of teachers.

The study will not resolve the differences in opinions about the merits of TVA models for teacher evaluation or the intended or unanticipated consequences of the application of TVA models. Rather, we aim to add evidence concerning the technical merits of TVA models to ongoing policy debates. Other studies have evaluated alternative TVA models using either actual or simulated data and contributed much to what we know about their strengths and weaknesses in using TVA models to estimate causal effects (Ballou, Wright & Sanders 2004, Guarino, Reckase & Wooldridge, forthcoming; Guarino, Reckase, Stacy, & Wooldridge 2013, Rothstein, 2010; Schochet & Chiang, 2010; McCaffrey, Lockwood, Koretz, Louis & Hamilton, 2004; Tekwe et al., 2004). The contribution of this study is that we use both actual and simulated data to test six commonly used TVA models on criteria relevant for teacher evaluation policy-making. We compare the performance of alternative TVA models using the same data and criteria. However, the comparisons are limited to models that provide estimates of individual teacher effectiveness and, therefore, do not include alternative means of evaluating teachers, such as principal ratings, ratings by independent observers, parent surveys, or student surveys. Thus, a major limitation of this study and other TVA model evaluations is that they compare how well these models perform but cannot provide direct comparisons about how these models compare to the other sources of data for evaluating teachers.

We now turn to a brief background on TVA estimates for individual teacher effectiveness. Second, we outline a causal theory upon which these estimates are based and the assumptions necessary to treat the estimates as unbiased estimates of the causal effects of individual teachers. Third, we define the six TVA models we assess. Fourth, we describe the actual and simulated data used for this study. Fifth, we discuss our findings and their implications in the final section of the paper.

BACKGROUND ON THE USE OF VALUE-ADDED MODELS FOR TEACHER EVALUATION

The use of value-added models for teacher evaluation is a logical extension of the movement towards standards-based education and accountability for outcomes that has come to dominate U.S. education policy in recent years. These policies have been manifest in federal legislation, No Child Left Behind (NCLB), and state standards, assessment, and accountability systems, several of which pre-dated NCLB. Significant research has been conducted on these policies (Carnoy & Loeb 2001; Hanushek and Raymond 2004; author; Dee and Jacob 2011; author) with mostly positive effects of school-based accountability on test score achievement. A legacy of these policies is the ubiquitous collection of state assessment data for reading and mathematics in 3rd through 8th grades as well as the creation of data infrastructures in most states that allow for linking student data, including current and prior test scores, to their teachers by subject. Linking students' past and present assessment scores to their teachers provides the minimum data necessary to generate estimates of individual teachers' effects on student test scores using TVA models.

This capacity was first made operational by William Sanders and associates, who in the mid-1990s began to link student assessment data to the students' teachers and estimate the ef-

fects of teachers who taught tested grades and subjects in Tennessee, in what became known as the Tennessee Value-Added Assessment System (TVAAS, now known as the Educational Value-Added Assessment System or EVAAS: Sanders, Saxton & Horn 1997; Ballou, Sanders, and Wright 2004; Amrein-Beardsley 2008). The most well-known EVAAS model uses a multiple membership, multiple classification (MMMC) random effects model (Browne, Goldstein & Rasbash, 2001) that requires only students' test scores and identification of the students' teachers, schools, and school districts by grade. While the data requirements are limited, the statistical and computational methods are very sophisticated and require substantial computational capacity (McCaffrey, et al. 2004). Many additional approaches have been developed in recent years, several of which will be described later in this section, to improve upon this value-added model.

More recently, school districts and states have begun to use TVA models for teacher evaluations (Scherrer 2011; NCTQ 2013). In addition, states were required to incorporate measures of student performance into teacher evaluations and teacher preparation program evaluations to successfully compete for federal Race to the Top (RttT) funds (author). The Los Angeles Times took the process further generating their own TVA scores for Los Angeles Unified School District teachers, releasing them on a searchable data base, and providing lists of top value-added teachers and schools and lowest value-added schools (LA Times n.d., downloaded from http://projects.latimes.com/value-added/). In February 2012, the New York Times followed suit and made teacher value-added scores for New York City teachers available as well. Clearly, the capacity to estimate individual teachers' effects has led to producing TVA estimates and it is unlikely to abate although it remains extremely controversial and some states are attempting to prohibit their pubic release.

CAUSAL ESTIMATES OF TEACHER EFFECTS ON STUDENT TEST SCORE CHANGE

In recent years, the potential outcomes framework, also known as Rubin's causal model, has provided educational researchers with a theory that identifies the assumptions needed for estimating causal effects (Holland 1986; Rubin 2005, 2008). Rubin, Stuart and Zanutto (2004) applied this framework to teacher evaluation and school accountability and Readon and Raudenbush (2009) explained the six assumptions needed to treat the estimates as causal effects. The centerpiece of the potential outcomes framework is counterfactual reasoning in which each teacher is considered a "treatment" and each student has a potential outcome under each possible teacher. Because only one treatment is realized and only one outcome per student for a given subject is observed, a challenge for estimating teacher effects is to identify an appropriate counterfactual. In TVA models the average treatment effect for a teacher (ATE_j) is the mean difference in the test score change of students assigned to teacher j minus the test score gains that would be expected for those students if assigned to the other possible teachers, often symbolized as .j (which is referred to as not j). An obvious candidate for operationalizing the counterfactual observation, .j, is the average level of teacher effectiveness, given that

the more generally accepted mechanism for creating a counterfactual, random assignment, is not feasible and perhaps not desirable for educational reasons (See Guarino et al. 2013, p 12).

Readon and Raudenbush (2009) described six assumptions that are needed to consider school effects and (by extension here) teacher effects as causal. Defining assumptions are (1) manipulability: each student must have the (theoretical) possibility of having an outcome under each teacher in the population and (2) the stable unit value treatment assumption (SUTVA): the potential outcome of each student is not affected by the assignment of other students to treatments. Estimating assumptions are (3) student test scores are interval scaled and (4) causal effects are homogeneous across participants in the same treatment group. Identifying assumptions are (5) the unconfounded assignment of students and teachers (also known as the strong ignorability of assignment to treatment) and (6) each teacher is assigned a group of "representatively heterogeneous" students such that there is sufficient overlap in the distribution of students assigned to each teacher to constitute "common support" for identifying causal effects.

While each of these assumptions has implications for TVA models, two of the assumptions have garnered more attention because they are particularly challenging for TVA model estimates to be considered causal: (1) the unconfounded assignment of students and teachers and (2) the absence of peer or classroom composition effects (SUTVA). Using actual data it is impossible to determine if alternative TVA models correctly estimate teachers' effects in the face of violations of the assumptions since teachers' true effects are unknown. However, "falsification tests" can be run to see if logically impossible results are obtained, which could indicate violations of assumptions are causing problems with the teacher effectiveness estimates. Rothstein (2010) ran such a test and found that 5th grade teachers appear to have had an effect on students' 4th grade test scores which is impossible if the TVA models were correctly estimating teacher effects and indicates that the models were not robust to violation of assumptions, likely unconfounded assignment. Several studies challenge Rothstein's findings. Koedel & Betts (2009) confirmed Rothstein's findings, though they also show that by adding more years of data, TVAs can be shown to be minimally biased. Chetty, Friedman & Rockoff (2011) suggest that Rothstein's test indicates only the possibility of unobserved variable bias, and in their examination of the effects of movements of effective teachers, demonstrate that the bias from mean teachers' value-added scores is minimal. Goldhaber & Chaplin (2012) also demonstrate that Rothstein's test suggests the possibility of unobserved variable bias from student/teacher sorting, but argue that Rothstein's test is flawed since it holds even in conditions where students and teachers are randomly assigned and, therefore, are not systematically sorted. Kane, McCaffrey, Miller, and Staiger (2013) added to the validation of TVA scores by randomly assigning students and teachers to classrooms, then estimating teachers' value-added scores, and comparing those estimates to estimates of the teachers' value-added scores when students and teachers were assigned to classrooms based on existing practice in six large school districts.

The unconfounded assignment assumption violation occurs as a result of failing to include the confounding variables in the TVA models that affect student test performance and are also associated with student-teacher assignment. Two types of violations of unconfounded assign-

ment may occur. Most commonly, positive assignment is hypothesized that involves assigning students with higher levels of personal motivation, academic ability, and parental support to teachers who are more effective, through tracking or parental pressure. Positive assignment would results in over-estimating the effectiveness of teachers. But compensatory assignment practices in which the most effective teachers are assigned students with the fewest parental or community resources in order to increase these students' chance for success could produce the opposite problem. Compensatory assignment could result in underestimating teachers' effectiveness to the extent that the missing inputs to student learning, in this example parental and community resources, are associated with student test performance and assignment of the students to teachers.

The SUTVA assumption rules out effects based on composition of the classroom, including those attributable to peer interactions and those between peers and teachers. These effects may be positive, say, when more learning occurs in a class with higher achieving peers from peer tutoring or students benefitting from other students' questions and the teacher's response to those questions, or negative, which can occur when students disrupt instruction or the teacher attends to certain students more than others. If unaddressed, this problem may result in teacher effect estimates that are either too large (positive) or too small (negative) thereby biasing the effect estimates in unpredictable ways.

It is important to understand that violating these assumptions does not automatically indicate that teacher effect estimates will be biased in any meaningful way. Many statistical models are robust to violations of certain assumptions but not others. For example, for the proof that ordinary least squares regression yields the best linear unbiased estimates of population parameters, the independent variables must be fixed but the assumption is routinely violated in research and regression is generally considered robust to the violation of that assumption. There is greater risk associated with these violations when the parameters of interest are not the averages in a population of teachers but the effectiveness of individual teachers. The question of interest for the present study is the extent to which TVA models generate individual teacher effectiveness estimates that are robust to the violation of the SUTVA or unconfounded assignment assumptions.

In this study we investigate robustness of alternative TVA models to violations of assumptions required for the estimates to be considered unbiased by the potential outcomes framework. In addition, we will raise questions of precision or reliability of effect estimates. For the purposes of teacher evaluation, we must assess the extent to which the teacher effect estimates are unbiased or recover the "true effects" of teacher and the extent to which they are sufficiently precise or reliable to produce stable estimates over time or identify highly effective or highly ineffective teachers. In order to cover the full range of performance issues relevant to TVA models, we use actual and simulated data that we describe in a later section.

DEFINITION OF TEACHER VALUE-ADDED MODELS

For the purposes of this study, we identified six TVA models that could be or in fact are being implemented with a dataset from a moderately large state: (1) a nested random effects model;

(2) student fixed effect model, (3) teacher fixed effects model; (4) dynamic ordinary least squares, (5) a hybrid fixed and random effects model; (6) a student growth percentile model. We describe the models in each category briefly below.

Nested Random Effects Models

In nested random effects models, which include hierarchical linear models (HLM) and are also referred to as multilevel models (Raudenbush & Bryk, 2002), the multiple levels of data—student, teacher, and school—are modeled as units nested within a hierarchy in which each lower level (e.g., student) is nested within one unique higher level unit. Because of this need for uniqueness in nesting, these models are usually implemented as single-year cross-sectional models such that re-sorting of students from year to year is not taken into account. Instead, prior test scores are included in the model to account for the effects of previous years' teachers on individual students' current performance. A general specification for these models is a three level model (HLM3) as follows:

$$(1) \qquad Y_{itsw} = X_{its}\beta_{ts} + X_s\beta_s + \beta_{w-1}Y_{i,w-1} + \beta_{w-2}Y_{i,w-2} + u_s + u_{ts} + e_{its}$$

Subscripts for the student (i), teacher (t), school (s) and period (w = current; w-1 = one year lag; w-2 = two year lag) are included. The variables specified in the model include: a test score on a standardized end-of-grade exam in a selected subject area in a specified period (Y); a vector of student, family or home characteristics that are associated with the accumulation of knowledge (X_{its}), a vector of school characteristics (X_s). Therefore, Y appears both as the dependent variable in the current period, as well as predictors of student achievement in the current period as a one-year lag ($Y_{i,w-1}$) and a two-year lag ($Y_{i,w-2}$). Errors for the school (u_s), the teacher (u_{ts}), and the student (e_{its}) are included. A constant term is also specified. An empirical Bayes' residual or empirical best linear unbiased predictor estimate of u_{ts} (the teacher random effect) defines the teacher effect. Use of the current test score as the dependent variable rather than a gain score specification is preferable since the gain specification assumes that there is no decay or forgetting on the part of a student from one year to the next and forces any decay into the student level error term.

Fixed effects models. The second major type of TVA models consists of student or teacher fixed effects that require panel data and can be estimated using standard econometric approaches for panel data to control for the error in the students' prior test scores. Student fixed effects produce estimates using only within-student variance. Teacher fixed effects models use indicators for each teacher to explicitly estimate an effect for each teacher. In fixed effects models in general, the estimates of the teacher effects are assumed to be adjusted for confounders that vary "between" the fixed units but not "within" (over time for student fixed effects, or over students for teacher fixed effects). We implemented two major variations of the fixed effects models, a student fixed effects model (SFE), and a teacher fixed effects model (TFE).

In the student fixed effect model (SFE) all years of a three-year panel were used, and all characteristics were de-meaned:

(2) $$(Y_{it} - \bar{Y}_i) = (\mu_{it} - \bar{\mu}_i) + (\alpha_i - \bar{\alpha}_i) + (e_i - \bar{e}_i)$$

Within-student means are represented by the terms with bars (e.g., $\bar{Y}_i$). The time-varying predictors of student achievement are represented by μ_{ti}; the student fixed effect by α_i, which takes up the fixed effects of time-invariant confounders, such as predispositions to learning or academic ability. Because these are invariant within student, the demeaning eliminates these variables from the model. Time-varying effects of these predispositions, if they exist, are not addressed in this model. The teacher effect is estimated as the mean of the residuals within each teacher $(e_i - \bar{e}_i)$.

The teacher fixed effects model (TFE) was implemented as a cross-sectional model focusing on the nesting of students in teachers and estimated using indicator variables for each teacher. One approach to the TFE is to de-mean at the teacher level, similar to how student fixed effects are de-meaned in the SFE. The demeaning eliminates the teacher effects, however, which is not the goal in the TFE. Instead, the teacher effects must be preserved. An equivalent approach to de-meaning is to use indicators for each teacher, and obtain the teacher effect estimates from the coefficients on these indicators:

(3) $$Y_{it} = X_{it}\beta_t + \beta Y_{i,w-1} + \alpha_t + e_{it}$$

The teacher fixed effect model is similar to the HLM3. Instead of estimating the teacher effect with the random effect u_{ts}, however, the teacher effect is estimated using the indicator variables represented by α_t. A difference between the teacher fixed effects and random effects is that the teachers' effects are directly accounted for in the model by the inclusion of the indicator variables. This could be a strength if the assignment of students to teachers is fully accounted for by the prior test scores but may be confounded if other variables that are excluded from the model are correlated to student achievement and teacher assignment. Another difference between the teacher fixed and random effects estimates is due to Bayesian shrinkage in accordance with the reliability of each teacher's sample mean (Wooldridge 2009).

Dynamic Ordinary Least Squares

A pooled regression model (labeled by Guarino et al. [2012] as a dynamic ordinary least squares, or DOLS), which uses the panel of data but ignores the nesting of time within students, was also estimated. In this model, each observation is treated as independent:

(4) $$Y_{itw} = X_{itw}\beta_t + \beta Y_{i,w-1} + e_{itw}$$

The DOLS bears a resemblance to both the HLM3 and TFE models, but uses the panel of data over multiple years instead of treating each grade level as a separate cross section. The teacher scores are estimated for all grade levels from teacher averages of the residual term e_{itw}.

Hybrid Fixed and Random Effects Models

The hybrid approach uses both random effects to estimate the teacher effect as an empirical Bayes residual or shrinkage estimate (Wright, White, Sanders & Horn, 2010) as well as fixed effects—usually implemented as unit-specific dummy variables—to control for confounders of the teacher effect. Accordingly, covariates are not generally included in the models (Ballou, Sanders & Wright, 2004). Two types of these models have been developed and put into use in several districts and states: (1) multivariate response model (MRM); (2) the univariate response model (URM). We tested the URM, which is the most flexible models and has been implemented on a statewide basis to estimate individual teacher scores for evaluation purposes.

The URM is random effects model that uses only students' previous test scores across multiple subjects and their teacher and school assignment (Wright, White, Sanders & Rivers, 2010). Fixed effects are incorporated via de-meaning of pretests (and not the dependent variable) using an extensive multi-step process as described in Wright, White, Sanders and Rivers (2010), and summarized here: a composite is calculated using the predicted deviations from the school mean of current student performance. These deviations from current student performance are estimated from the deviations of multiple pretests (2 periods of math and 2 periods of reading, for example) from their own school means. This composite is then entered into a nested random effects model as the only covariate:

$$y_{it} = \beta_0 + \beta_1 C_i + u_t + e_{it} \quad (5)$$

The nesting in this final model is of students within teachers in one school year with no accounting for the nesting within schools; in addition the teacher effect estimation uses only one subject, despite the use of two subjects' data to estimate the composite test score.

Student Growth Percentile

The five models described in the previous sections are based on changes in student performance on achievement tests and are interval-scaled approaches. That is, the TVA estimate is based on an underlying metric intended to represent learning relative to some absolute standard. A sixth model, the student growth percentile model (SGP; Betebenner, 2011) bases the TVA estimate on students' relative position among peers who have demonstrated similar performance in the past. To estimate TVAs using SGP, a quantile regression is estimated first, with a prediction model for every percentile of student growth, as follows, with two previous years' achievement as predictors:

$$y_{it} = \beta_0 + \beta_{w-1} Y_{i,w-1} + \beta_{w-2} Y_{i,w-2} + e_{it} \quad (6)$$

Actual student performance is compared to predicted performance within each percentile, identifying each student's growth percentile. The probit function converts each student's percentile to a normal curve equivalent (NCE), which has the same distributional form as a z score. The mean of the NCE within each teacher gives the TVA estimate. Student percentile growth

models do not try to estimate the magnitude of the effects of individual teachers but rather they rank order teachers in terms of their students' actual versus predicted growth. They have been included here since they are currently being used for estimating the effectiveness of individual teachers in eight states and the District of Columbia (Walsh & Isenberg 2013).

Summary of models. A summary description of the six TVA models used in this study is presented in Table 1. All of the models except for two (the TFE and SGP) obtains the teacher effect indirectly via a type of residual (the TFE estimates the teachers effects directly from the coefficients on the teacher indicator variables and the SGP uses the students' average percentile growth averaged for each teacher). All TVA models include prior student test scores to address confounded assignment. In addition, the random effects models, as well as the teacher fixed effect model, attempt to control for confounded assignment through covariates, available or measured inputs to learning or proxies for inputs to learning that may be highly correlated with the actual inputs. None of the models explicitly addresses SUTVA, though some of the models may make accommodations for violations of these assumptions. For example, the random effects models may include classroom level covariates, which could distinguish between effects due to the teacher and effects due to the teacher's interaction with each group of pupils to which they are assigned.

Table 1: Summary of Teacher Value-Added Models

	Cross-Sectional or Panel	Time Invariant Covariates	Lagged Outcomes (Pretests)	Teacher Effect Parameter	School Random Effect
HLM3	Cross-sectional	Yes	2 in each of 2 subjects	Teacher random effect (EB shrinkage estimator)	Yes
SFE	Panel	Differenced to zero	None (all outcomes as demeaned DV)	Mean of within-teacher residuals	N/A
TFE	Cross-sectional	Yes	1 in same subject	Teacher fixed effect (dummy variable)	
DOLS	Panel (2 periods)	Yes	1 in same subject	Mean of within-teacher residuals	
URM (EVAAS)	Cross-sectional	No	2 in each of 2 subjects used to calculate composite	Teacher random effect (EB shrinkage estimator)	No
SGP	Panel	No	2 in same subject	Teacher mean of probit-transformed percentiles of growth	N/A

STUDY DATA AND METHODS

The aim of this study is to provide an assessment of six TVA models to be used in teacher evaluations. We propose three overarching criteria that should be used to judge TVA models:

1. TVA models should reproduce teachers' "true effects" and minimize total error when SUTVA (assumption of no classroom effect on students' test score change) or the assumption of unconfounded student-teacher is violated. *Justification:* TVA models should be ro-

bust to common violations of assumptions needed to consider teachers' scores as causal estimates.

2. TVA models should not falsely identify average teachers as highly ineffective when the SUTVA or unconfounded assignment assumptions are violated.[2] *Justification:* Many proponents have advocated that teachers TVA scores be used for high stakes consequences including denial of tenure. Falsely identifying teachers who are not ineffective as ineffective could cause the loss of effective teachers and correspondingly, because for every effective teacher identified as ineffective, an ineffective teacher is found to be effective, false identification misses opportunities to aid or remove truly ineffective teachers.

3. TVA scores should be reasonably reliable from one year to the next; that is, the change in teachers' relative performance from year-to-year should be moderate. *Justification:* Proponents have argued that rewards for high TVA scores could be used as incentives for higher teacher performance. However, if the scores change in seemingly unpredictable ways or exhibit large fluctuations, any motivational effects of the rewards will be largely undermined.

In order to assess TVA models on each of these criteria we employed both simulated and actual data.

Simulated data. Simulated data must be used when researchers are assessing the extent to which TVA models can recover the "true effects of teachers" since the "true effects of teachers" cannot be directly measured or otherwise obtained in actual data, except in the rare situation of random assignment of teachers and students to classes. The simulated data were calibrated using actual North Carolina data on elementary and middle schools. Specifically, variance in North Carolina student test scores was decomposed into student, teacher, and school levels, showing that approximately 13% of the variance was at the teacher level in math and 9% in reading. These calibrations are consistent with findings from other studies that look at classroom-level rather than teacher-level variance such as Nye, Konstantopolous and Hedges (2004), which found a range of values between 10 and 14% variance at the classroom level. Data simulations always require simplifications from the data generated in actual circumstances but can in many ways be generated to mimic the characteristics of actual data. In this case, we did not allow for either teachers or students to change schools; we estimated effects for one elementary grade, 5th, and one middle school grade, 8th; complete data (no missing data) were simulated for all students and teachers; and the dataset consisted of multiple districts but was not as large as the actual data we used from North Carolina. To ensure the disturbances in the TVA estimates reflected violations of confounding or SUTVA in the contemporaneous setting and not violations in earlier periods, we omitted persistence of teacher effects from earlier periods (See McCaffrey et al. 2004 for an explanation of persistence effects in value-added models). This decision was subjected to a robustness analysis. Two data generation processes, both based on the extant literature examining value-added models, were used: (1) variance decomposition, similar to that used in Schochet & Chiang (2011); (2) heterogene-

[2] While it is important to identify both highly effective and highly ineffective teachers, we present only the analysis of highly ineffective teachers because the consequences for these teachers can be more severe and the findings for ineffective and effective teachers are very similar.

ous teacher effects, similar to that used in Guarino, Reckase & Wooldridge (2012). We used the variance decomposition data generating process to test the impact of violations of SUTVA, since it allows us to add classroom effects that are independent of teacher effects. We used the heterogeneous teacher effects data generating process to test the impact of violations of confoundedness, since it allows us to add a correlation between teacher and student effects that could result from assignment of students to teachers. As far as we can determine, this represents the first study that allows a comparison of the impacts of the violation of these two assumptions.

For the variance decomposition simulation, which tests the impact of a violation of SUTVA, each student's test score was partitioned into six components: student, classroom, teacher, and school, average state effect, and a random error. The teacher effect was homogeneous for all students taught by each teacher. Closely following Schochet and Chaing (2010), the classroom effect was set at 4% of the total variance to simulate the violation of the SUTVA assumption and at 0 to represent the absence of the violation of SUTVA. The statewide mean μ_{jw} for each of six standard normal $\sim N(0,1)$ test scores was specified and then added to time-invariant effects within each subject area for student (ϕ_{ik}), classroom (ϕ_{ck}), teacher (ϕ_{jk}), school (ϕ_{sk}), and district (ϕ_{dk}) effects created via the variance decomposition, as well as a random or measurement error component ($r_{icjsdkw}$), to arrive at the total score for each student in each grade level and subject, as follows (with i = student, c = classroom, j = teacher, s = school, and k = subject, all defined as above, adding d for district):

$$(8) \qquad Y_{icjsdkw} = \mu_{jw} + \phi_{ik} + \phi_{ck} + \phi_{jk} + \phi_{sk} + \phi_{dk} + r_{icjsdkw}$$

The true teacher effect was the subject-area specific teacher input to student learning entered into equation 8 (ϕ_{jk}). Teachers were randomly assigned between 17 and 23 students per year and assumed to teach both reading/language arts and math in elementary grades. For any single elementary cohort of students when teachers and students are primarily assigned to the same classroom for the entire day, the teacher effect cannot be disentangled from the classroom effect, therefore we generated data for two cohorts of students. In each simulated cohort, there were 99,252 records generated, consisting of 16,542 students taking three end-of-grade exams in each of two subjects. A total of 833 teachers were simulated across 184 schools in 14 school districts. For the middle school simulation, we assumed that each teacher taught five classes of either mathematics or reading/language arts implying that more students would be needed within any cohort in order to ensure a sufficiently large teacher sample than in the elementary school simulation, but that only one cohort would suffice to separately identify classroom and teacher effects during VAM estimation. In the one simulated cohort, there were 241,284 total records, consisting of 40,214 students each enrolled in three grade levels over two subjects. A total of 410 teachers in 154 schools in 30 districts were simulated.

For the heterogeneous effects simulation, which tests the impact of a violation of unconfounded student assignment, students, classrooms, teachers and schools covariates were included in the data generation process but omitted from the TVA estimates, thus producing the effects that would be expected to occur in positive assignment (positive correlations) and neg-

ative assignment (negative correlations). The correlated covariates included one time-invariant student background effect for each of two subjects (μ_{ik}); one classroom effect for each grade level (of 3) for a specific subject (μ_{ckw}, with w = 1, 2, 3); one teacher effect for each grade level for a specific subject (μ_{ikw}); and one grade- and subject-invariant school effect (μ_s; the school effect subsumed the district level effect). Assignment was simulated using a correlation of .20 (positive assignment) and -0.20 (negative assignment) between student effects and the effects for classroom, teacher, and school across 3 grade levels. Correlations of -0.20 and 0.20 are consistent with medium effect sizes in education research. The classroom, teacher and school effects also had invariant correlations with each other (.20). Over the three grade levels, teachers were correlated at .50. All of these inputs were calibrated to actual North Carolina data.

To be as realistic as possible, parameter variance was assigned to each student, classroom, teacher and school effect. These included $\eta_{ikw}$for student, $\eta_{ck}$for classroom, η_{kw} for teacher, and η_{sk} for school. A subject and grade-specific state grand mean (μ_{kw}) and residual ($e_{icgsdkw}$) were also estimated. All fixed and random effects were summed to produce the total achievement score for each student, as follows:

(9) $$Yicjskw = \mu_{kw} + \mu_{ikw} + \mu_{ckw} + \mu_{jkw} + \mu_s + \eta_{ikw} + \eta_{ck} + \eta_{jk} + \eta_{sk} + e_{icgsdkw}$$

The teacher effect, for the purposes of identifying a true value and estimating teacher effects from each VAM, was the teacher-level mean of the heterogeneous fixed teacher effect μ_{jkw}. For the elementary schools, we simulated 40,000 students in 2,000 classrooms, with 2 classrooms per teacher (representing 2 cohorts of students) and 1,000 teachers. For the middle grades, we simulated 200,000 students in 10,000 classrooms representing 10 classrooms per teacher in middle school (5 classrooms per teacher over 2 cohorts) and 1,000 teachers. Models were run with 100 simulations after conducting a sensitivity test that showed findings for versions with 1,000 simulations were nearly identical to those with 100 simulations.

Actual North Carolina Data. The analysis of actual data was conducted on data collected in North Carolina between 2007-08 and 2009-10, with some student test score data also available from 2005-06 and 2006-07. Both math and reading end-of-grade standardized exam scores were used. These data were used to examine the year-to-year reliability of actual estimates of teacher effectiveness as well as to estimate the number of teachers that the TVA models would identify as highly effective or ineffective. In addition to the lagged scores or pretests as specified in the model, all available student, peer and school characteristics were incorporated as covariates in the analysis when the TVA models allow them and incorporate them in standard practice. These included student race/ethnicity, subsidized lunch, limited English proficiency, disability and academic giftedness, within and between-year movement, under and over age indicators, peers' previous average standardized exam score, and an indicator of the year. Classroom composition covariates included an indicator that the classroom average was above the 75th percentile in limited English proficiency, disability or giftedness, free lunch eligibility, and overage students. School covariates included proportion by race/ethnicity, total per-pupil expenditures, percent subsidized lunch eligible, violent acts per 1,000 and suspension rates in previous year, and enrollment. No teacher characteristics were included in the data because

these teacher characteristics could explain a portion of the teacher effect that we actually want to estimate. The data used in this study consisted of all student records with valid test scores in 5th or 8th grade in North Carolina public schools during three focal years.

For the year-to-year consistency analysis, which required two sequential within-year estimates for each grade level, there were limitations to the amount of information available for the models that required multi-year panels to estimate (SFE and DOLS) for which four consecutive years of complete (test score and covariate) student would be required. Because only test score data were available prior to 2007-08, no covariates could be included in these models. There were 503,370 student records in 5th grade math (8,826 teachers); 298,323 student records in 8th grade math (3,655 teachers); 728,008 student records in 5th grade reading (9,402 teachers); and 298,323 student records in 8th grade reading (4,169 teachers). As far as we know based on the literature (e.g., Koedel & Betts, 2011; Sass, 2008; Goldhaber & Hansen, 2008) this is the first study to examine the year-to-year consistency of TVA model scores for such a wide range of nested, fixed and hybrid models.

FINDINGS

(1) Which TVA estimates are most highly correlated with the "true" effects or produce the minimum error?

In the absence of violating assumptions, the TVA scores are correlated between 0.96 and 0.91 for elementary grades and 0.93 and 0.87 for middle grades with the "true" effects in the simulated data (Table 2). Spearman's rank order correlations are used since TVA estimates may be used to rank teachers within a unit (state, district, or school). When the SUTVA assumption was violated with a classroom level effect of 4% of the variance, all correlations are reduced and the rank ordered performance of the models remains stable. Estimates from four models were more highly correlated with the teachers' true effects (HLM3, URM, SFE, and SGP) than the TFE or DOLS. The performance of the TVA models varied much more when unconfounded assignment was violated (ranges of 0.114 elementary grades to 0.077 middle grades) than when SUTVA was violated (range of 0.042 to 0.061). While the absolute performance of the TVA models varied substantially, with one model consistently the best performer across grade level and assumptions violated (HLM3; at or above .77 in all scenarios), which was likely due to the inclusion of covariates in this model. The URM appears to be the second best performer but did not perform as well when unconfoundedness was violated. All models performed better for the middle grades than elementary grades which could be expected from the increased size of the sample for each teacher. TFE and to a slightly lesser extent DOLS did much better when confoundedness was violated than when SUTVA was violated. TFE, DOLS and SFE all performed worst in one of the tests.

Table 2. Spearman Rank Order Correlations between TVA Teacher Effectiveness Estimates and "True Effect" by No Assumption Violation, SUTVA Violation, and Unconfoundedness Violation

Elementary School				**Middle School**		
	No Violation	SUTVA Violated (4% Classroom Variance)	Unconfoundedness Violated (ρ = .20)[1]	No Violation	SUTVA Violated (4% Classroom Variance)	Unconfoundedness Violated (ρ = .20)[1]
HLM3	0.955	0.864	0.772	0.932	0.912	0.822
SFE	0.941	0.851	0.646	0.914	0.893	0.729
TFE	0.909	0.822	0.712	0.869	0.851	0.769
DOLS	0.909	0.822	0.702	0.874	0.853	0.761
URM	0.946	0.856	0.688	0.903	0.901	0.769
SGP	0.942	0.853	0.663	0.917	0.897	0.745

[1]Correlation shown is between student and classroom, teacher, and school effects; correlation between classrooms or classroom and teacher is .20; correlation between teachers is .50; correlation between classroom or teacher and school is .20.

To test the total minimum error, we calculate the mean square error for each of the six TVA models (see table 3). When neither assumption was violated, the top three performing models are HLM, URM and SFE. The SGP was the worst performer with MSE of 0.617 in this scenario, most likely because it does not attempt to isolate the teachers' contribution or that of classrooms or schools to the student percentile growth (SPG calculates the total actual versus expected growth for the students in each teacher's class and attributes the entire difference to the teacher). The performance of all of the models degrades when SUTVA was violated and further degrades when unconfoundedness was violated, except for SGP, which performs better but not among the top in the latter comparisons. Across all of the violations of assumptions, the best performing model was HLM. SFE and URM perform slightly worse than the HLM model and the performance of the other models was somewhat erratic but always below these three models.

Table 3. Mean Square Error (Estimated Minus "True" Effect Squared) by No Assumption Violation, SUTVA Violation, and Unconfoundedness Violation

	Elementary School			**Middle School**		
	No Violation	SUTVA Violated (4% Classroom Variance)	Unconfoundedness Violated (ρ = .20)[1]	No Violation	SUTVA Violated (4% Classroom Variance)	Unconfoundedness Violated (ρ = .20)[1]
HLM3	0.011	0.025	0.054	0.016	0.015	0.040
SFE	0.021	0.025	0.088	0.025	0.020	0.059
TFE	0.025	0.041	0.125	0.040	0.034	0.081
DOLS	0.025	0.041	0.132	0.038	0.033	0.084
URM	0.014	0.030	0.080	0.022	0.020	0.060
SGP	0.617	0.620	0.100	0.692	0.531	0.066

[1]Correlation shown is between student and classroom, teacher, and school effects; correlation between classrooms or classroom and teacher is .20; correlation between teachers is .50; correlation between classroom or teacher and school is .20.

(2) How many "false negatives" or teachers who are not in the lowest 20 percent do TVA models identify as in the lowest 20 percent?

The teachers identified as ineffective are those for whom some consequences including dismissal or required professional development may be initiated by the evaluation system and teachers labeled false negatives would be those who are identified and receive the consequences of being ineffective but actually performed well enough to avoid them. We now test for "false negatives" using a criterion to identify the lowest performing 20 percent of the teacher workforce or those teachers with TVA scores more than 0.842 standard deviation units below the mean whose "true" effectiveness is greater than -0.842 standard deviation units.

In the absence of assumption violations, all six TVA models would falsely identify approximately 3-5 percent of the teachers in elementary grades and middle grades as ineffective as shown in Table 4. In other words, when the policy goal is to identify the 20 percent of the teaching force that is the lowest performing, these models would identify 15-17 percent correctly, meaning that 3-5 percent of those identified as low performing should not be and 3-5 percent of those who should be were not identified. In a state the approximately the size of North Carolina (assuming 9000 5^{th} grade teachers and 4000 8^{th} grade mathematics or reading/language arts teachers), TVA models would falsely identify between 284 and 407 5^{th} grade teachers and between 152 and 216 8^{th} grade math or reading/language arts teacher as ineffective and an equal number who were ineffective were not identified. The average true z-score for these falsely identified teachers' effectiveness was between -0.49 standard deviation units (SDU) and -0.43 SDU for elementary teachers and between -0.57 and -0.46 SDU for middle school teachers, which indicates that on average, the estimated scores are worse (at -0.842 or lower) than their true effect estimates. For the best performing model—HLM3 – the average performance of a teacher falsely identified as ineffective was approximately -0.6, which is more than one-half standard deviation unit below the mean and closer to the cutoff than in the other models. Four of the models perform similarly but two, TFE and DOLS are somewhat worse in the absence of assumption violations.

The performance of the TVA models was reduced by similar amounts when the SUTVA assumption was violated. The TVA models would falsely identify approximately 4 to 7 percent of elementary and middle school teachers as ineffective, which would amount to approximately 500-600 elementary school teachers and about 200 middle school teachers in each subject using the same assumptions about the number of teachers per grade as above. Model performance was more consistent than in the scenario with no violation, though the HLM3 was still the best performing model (along with the URM for the elementary level). When the assumption of unconfounded assignment was violated by positive assignment, TVA models performance varies considerably with the best performing model, HLM3, falsely identifying 7.3% or 658 elementary teachers and 6.9% of middle school teachers, for a total of 276; and the worst performing model, SFE, falsely identifying 9.3% (835) of the 5^{th} grade teachers and 7.9% (317) of the 8^{th} grade teachers as ineffective. None of the other four models were consistently better or worse performers.

Table 4. Teachers Falsely Identified as Ineffective, Simulated Data, Unconfoundedness Violation

Cutoff = Z = -0.842/20% of Teachers Ineffective									
	No Violation			SUTVA Violation (4% Classroom Variance)			Unconfoundedness Violation (ρ = .20)[2]		
	%	No. of Teachers Affected[1]	Mean True Z Score of Affected Teachers	%	No. of Teachers Affected[1]	Mean True Z Score of Affected Teachers	%	No. of Teachers Affected[1]	Mean True Z Score of Affected Teachers
Elementary School									
HLM3	3.2%	284	-0.64	5.7%	510	-0.49	7.4%	668	-0.37
SFE	3.6%	322	-0.61	5.9%	531	-0.46	9.3%	835	-0.24
TFE	4.5%	407	-0.55	6.5%	588	-0.43	8.4%	756	-0.30
DOLS	4.5%	402	-0.55	6.5%	588	-0.43	8.6%	773	-0.30
URM	3.4%	308	-0.63	5.7%	517	-0.47	8.8%	789	-0.28
SGP	3.5%	313	-0.62	5.9%	527	-0.47	9.0%	811	-0.26
Middle School									
HLM3	3.8%	152	-0.60	4.4%	176	-0.57	6.4%	256	-0.43
SFE	4.5%	180	-0.57	5.0%	199	-0.53	7.9%	317	-0.31
TFE	5.4%	216	-0.49	5.9%	236	-0.46	7.5%	299	-0.36
DOLS	5.3%	212	-0.50	5.9%	234	-0.47	7.4%	296	-0.35
URM	4.3%	172	-0.58	4.8%	191	-0.55	7.3%	292	-0.36
SGP	4.4%	176	-0.57	4.9%	195	-0.54	7.7%	308	-0.33

[1]Assuming 9,000 5th grade teachers; 4,000 8th grade teachers.

[2]Correlation (ρ) shown is between student and classroom, teacher and school; correlation between classrooms or classroom and teacher is .20; correlation between teachers is .50; correlation between classroom or teacher and school is .20.

(3) How many teachers are in the same performance quintile from one-year to the next and how many switch from highjest to lowest performance quintile or vice versa from one year to the next?

Instability in classifying the performance of teachers from one year to the next can undermine the credibility of TVA model scores and their utility as a criterion for incentives, such as performance bonuses. For this test, we used the actual North Carolina data described above. In table 5, we present our test of year-to-year reliability, the percentage of teachers in the same performance quintile from one year to the next and a test of unreliability, the percentage of teachers that switch from top to bottom quintile or vice versa from one year to the next. TVA model performance varies considerably but perhaps more striking is that the TVA model performances on these tests vary substantially by subject and grade. For example, the year-to-year consistency of estimates for the high performing DOLS models ranges from 53.9% for 8th grade math to 39.2 percent for 5th grade reading, probably due to pooling the data from two years. In terms of overall consistency, the higher performing models (same quintile both years) are the DOLS (in both math and reading across grade level); the next best models varied depending on grade level and subject, but included the TFE (reading, 5th grade), SGP (math, 8th grade), URM (math, 5th grade), and TFE (reading, 8th grade).

Table 5. Consistency of Rankings Over Consecutive Years

Quintiles of Teachers				
Elementary School (5th Grade)				
	Same Quintile Both Years		Switch from Lowest to Highest or Highest to Lowest	
	Math	Reading	Math	Reading
HLM3	30.0	24.5	3.2	5.9
SFE	34.6	27.5	2.2	4.3
TFE	33.3	28.9	1.7	3.9
DOLS	44.5	39.2	0.2	0.8
URM	35.1	28.3	1.7	4.3
SGP	34.5	28.3	2.1	3.9
Middle School (8th Grade)				
	Same Quintile Both Years		Switch from Lowest to Highest or Highest to Lowest	
	Math	Reading	Math	Reading
HLM3	34.6	26.4	1.3	7.7
SFE	37.3	31.2	1.8	4.2
TFE	41.0	34.5	1.3	3.3
DOLS	53.9	51.0	0.4	0.8
URM	39.8	32.4	0.4	3.0
SGP	42.9	34.3	1.5	4.4

In terms of extreme switchers, the DOLS was again the top model, with under 1% in this category; the next best model was URM. In contrast to the findings of the previous criteria, the HLM3 was a lower performing model on both same quintile and extreme switchers.

ROBUSTNESS

An assumption of the data generation process—zero persistence in teacher effectiveness across time—was subjected to robustness diagnostics. This assumption focuses attention only on the contemporaneous inputs to student learning, including the teacher. Although this helps with the purpose of the current study it is not realistic. Therefore, we repeated the tests using data in which effects of prior inputs persisted in current effects. On the Spearman rank order, the performance of all models except the HLM3 was degraded under at least one scenario. For example, the SFE fell from .851 to .728 under a 4% SUTVA violation. However, the HLM3 was fairly consistent across models (steady at 0.864). Performance of the URM which assumes persistence with no decay was also fairly consistent under a SUTVA violation (steady at 0.856) but less so under an unconfoundedness violation (from 0.688 without persistence to 0.593 with persistence). On falsely identifying teachers as ineffective, degradation of performance was not substantial for any models under the SUTVA violations, but for the unconfoundedness violation, only the HLM3 performed similarly.

DISCUSSION

The findings presented above remind us of the statement attributed to the famous statistician George E. P. Box, "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful" (Box and Draper, 1987, p. 74). All of the TVA models make errors and have a degree of inconsistency in estimating the effectiveness of individual teachers of tested grades and subjects. We believe that three questions must be answered to make the decisions required to incorporate TVA model estimates into teachers' evaluations.

First, we should consider whether in the absolute or relative sense are TVA model scores good enough to include in the evaluations of individual teachers? When assumptions are violated the best TVA model scores rank order correlation with the "true" effect was between .76 - .89. In an absolute sense, to judge TVA models as adequate for use in teacher evaluation, one would have to accept the rank order correlations with "true" effects are within this range when assumptions are violated as we know they are likely to be. In addition, we would have to accept falsely identifying between 3.2 and 7.4 percent of the teachers as among the 20 percent of the poorest performers using the best performing model, HLM3. Another model, SFE, incorrectly identifies between 8 and 9 percent of the teachers as among the 20 percent of the poorest performers when the unconfoundedness assumption is violated and other models incorrectly identify between 7 and 9 percent of the poorest performing 20 percent.

In practice, no classification system will be perfect. Moreover, the TVA models may be much better at identifying poor performing teachers than any other means that are currently available. However, an implication of this analysis is that in the identification of the lowest performing teachers the false negatives have an equal number of those who did perform poorly but were not identified as such. Also, in the identification of highest performing teachers, the false positives will be approximately equal to the percentage of false negatives and another equal-sized group who should have been identified as high performing were not. In total the misclassified teachers (effective as ineffective + effective as highly effective + ineffective as effective + highly effective as effective) is approximately four times as large, which could mean that nearly 36 percent of the elementary grade teacher workforce may be incorrectly identified (about 30 percent for middle grade teachers). But the gains from identifying and intervening with lowest performers and encouraging high performers may outweigh the costs of misclassification depending on the interventions that are selected.

This does seem to make it clear that an implication of these findings about the performance of TVA models for teacher evaluation is that the advisability of their use depends on the consequences of the evaluation process that are to be meted out. For low stakes purposes, such as making professional development recommendations, the error rates seem to be reasonable; but, for high stakes purposes including sanctions such as denial of tenure, the rates of false negatives and positives seem to be too high. Although recent research shows that identifying teachers in the lowest performing group for two consecutive years would greatly reduce the percentage of teachers identified and logically the percentage of teachers misidentified (Winters & Cowen, 2013). In the middle of these high and low stakes purposes, a medium stakes purpose such as identifying less effective teachers for more observations, feedback and coach-

ing would seem to be reasonable. The latter would allow concentrating resources on teachers who are likely to most need them (even the effective teachers incorrectly identified as ineffective were on average substantially below average in the better performing models) and avoiding making errors that could do permanent harm, such as denial tenure for relatively effective teachers. As TVAs themselves do not identify teachers' strengths or weaknesses, follow-up observations may highlight deficits that are opportunities for professional development. Our role is not to make a judgment, absolute or relative, but to point out the potential for error in order to encourage informed judgments.

Second, which TVA models are sufficiently accurate for use in teacher evaluations? The answer to us seems that HLM3, the three-level hierarchical linear model with two years of pretest scores is the best performer overall in these tests, but this model performed poorly in the year to year consistency tests. The URM, SFE and DOLS models all performed reasonably well in certain tests and all performed better in terms of year-to-year consistency than HLM3. Again, model preference may vary based on the purpose of the evaluation and the degree to which those choosing a model believe that one assumption violation is more likely or more harmful. For example, as this study and prior studies show the DOLs models perform well if unconfoundedness is violated (Guarino, et al. forthcoming) so if violating unconfoundedness poses the greatest perceived risk, DOLS could be a reasonable choice. If an intended use of the teachers' scores is to determine eligibility for performance bonuses as an incentive for higher performance, the use of HLM3 could be questionable. But if the estimates are to be used to classify the lowest performing teachers, HLM3 seems to be a preferred choice.

Third, we believe that the year-to-year inconsistencies seem to beg the question, how much does teacher performance actually change from year-to-year? While extreme switchers – those who switch from top to bottom performance quintiles or vice versa—do emerge from all TVA models, the absolute rates are quite low with most models. Perhaps, more troubling is the lack of stability in the quintile categories of performance from year-to-year. We do not have independent information that could provide a guide for how teacher performance actually fluctuates from year-to-year. It may be that novice teachers who are gaining experience, teachers who switch schools or grades, and anomalous events are related to the observed year-to-year fluctuations. Understanding the extent of instability and the causes for the instability should be the subject of more research so that we can begin to isolate actual changes form year-to-year from model error.

Finally, we return to our original question in the title: "Are value-added models good enough for teacher evaluations." In truth, evidence alone will never be sufficient to answer this question and in practice, tolerance for error is likely to vary greatly among those who make decisions about and those who are affected by the TVA scores. With the increase in availability of the data needed for estimating TVA, the publication of TVA scores by large media outlets, and the use of these models by state and local education agencies to estimate individual teachers' effectiveness scores and make them available for the public or supervisors to use for evaluative purposes, it seems unlikely that the use of teachers' TVA scores will abate in the foreseeable future. The tests of the TVA models' performance in the face of assumption violations that are likely to occur in actual school settings should provoke caution in using the estimates for

teacher evaluation. However, in the absence of information about the performance of other methods of obtaining data on teachers' performance, it seems that the risks for incorporating TVA scores into teachers' evaluation is not excessive, especially when other measures are used for low to medium stakes purposes. However, the findings suggest to us that the use of TVA scores for high stakes decisions may be quite risky in terms of the direct consequences such as denying tenure to effective teachers not to mention potential negative side-effects such as deterring talented young people from pursuing teaching careers. Certainly, the use of multiple measures and perhaps, requiring poor performance in consecutive years would be more prudent that relying on a single or even combining two rounds of TVA scores (Winters & Cowen, 2013). Research should continue on the validity, reliability and consistency of both TVA models and alternative means of obtaining evaluative information on teachers' performance to reduce risk from using imperfect measures of teachers' performance and support policies that maximize the availability of high performing teachers.

REFERENCES

Amrein-Beardsley A. (2008). Methodological Concerns About the Education Value-Added Assessment System. Educational Researcher, 37 (2) 65-75.

Amrein-Beardsley A. & Collins, C. (2012).The SAS Education Value-Added Assessment System (SAS ® EVAAS ®) in the Houston Independent School District (HISD): Intended and unintended consequences. Education Policy Analysis Archives, 20 (12). Retrieved August 13, 2012, from http://epaa.asu.edu/ojs/article/view/1096.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' Rule to Estimate The Effect of Class Size on Scholastic Achievement*. *Quarterly Journal of Economics, 114*(2), 533-575.

Arellano, M. & Bond, S. (1991) Some Tests of Specification of Panel Data: Monte Carlo Evidence and an Application to Employment Equations. The Review of Economic Studies, 58, 277-298.

Baker, E. L. Barton, P. E. Darling-Hammond, L., Haertel, E., Ladd, H. E., Linn, R. L., Ravitch, D. Rothstein, R., Shavelson, R. L. & Shepard, L. A. (2010) Problems with the use of student test scores to evaluate teachers. Washington DC: Education Policy Institute.

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.

Box, G. E. P. & Draper, N. R. 1987. Empirical model-building and response surfaces. Oxford, England: John Wiley & Sons.

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2006. How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy* 1(2): 176-216.

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2009. Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis* 31(4): 416-440.

Boozer, M., & Rouse, C. (2001). Intraschool Variation in Class Size: Patterns and Implications. *Journal of Urban Economics, 50*(1), 163-189.

Browne, W. J., Goldstein, H. & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) Models. *Statistical Modeling 2001 (1)*: 103-124.

Carnoy, M., & Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2012). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper. No. 17699. Cambridge, MA: National Bureau of Economic Research.

Clotfelter, Charles, Helen Ladd, and Jacob Vigdor. 2007. Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review* 26(6): 673-682.

Clotfelter, Charles, Helen Ladd, and Jacob Vigdor. 2010. Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *The Journal of Human Resources* 45(3): 655-681.

Dee, T. S. & Jacob, B. 2011. The impact of No Child Left Behind on student achievement. Journal of Policy Analysis and Management 30 (3) 418-446.

Finn, J.D. & Achilles, C. M. 1990. Answers and Questions About Class Size: A Statewide Experiment. American Education Research Journal 27 (3) 557-577.

Ferguson, R. F. & Ladd H. F. 1996 How and why money matters: An analysis of Alabama schools in Ladd, H. F. (ed.) Holding school accountable: Performance based reform in education. Washington DC: Brookings Institution.

Goldhaber, D. & Chaplin, D. (2012). Assessing the "Rothstein Test": Does it really show that value-added models are biased? National Center for Analysis of Longitudinal Data in Education Research.

Goldhaber D. & Hansen, M. (2008). Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions. National Center for Analysis of Longitudinal Data in Education Research.

Gordon, R. Kane, T. J. and Staiger, D. O. (2006) Identifying effective teachers using performance on the job. Hamilton Project Discussion Paper. Washington DC: Brookings Institution.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (forthcoming). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*.

Guarino, C. M., Reckase, M. D., Stacy, B., & Wooldridge, J. M. (2013). A comparison of growth percentile and value-added models of teacher performance. IES working paper, Washington, DC.

Hanushek, E. A. (1994). An Exchange: Part II: Money Might Matter Somewhere: A Response to Hedges, Laine, and Greenwald. *Educational Researcher, 23*(4), 5-8.

Hanushek, E. A. (1999). Some Findings From an Independent Investigation of the Tennessee STAR Experiment and From Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis, 21*(2), 143-163.

Hanushek, E. A., & Raymond, M. E. (2004). The Effect of School Accountability Systems on the Level and Distribution of Student Achievement. *Journal of the European Economic Association, 2*(2-3), 406-415.

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. Journal of Policy Analysis and Management, 28(4), 693-699.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). An Exchange: Part I: Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes. *Educational Researcher, 23*(3), 5-14.

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. Journal of Policy Analysis and Management, 28(4), 700-712.

Holland, Paul W. (1986) Statistics and causal inference. *Journal of the American Statistical Association, 81*: 945-960.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615-631.

Koedel, C. &Betts, J. R. (2009). 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.

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions*. *Quarterly Journal of Economics, 114*(2), 497-532.

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.

National Council on Teacher Quality (2013). *State Policy Dashboard* downloaded March 9, 2014 at http://www.nctq.org/statePolicy/statePolicyIssues.do

Raudenbush, S. W. and A. S. Bryk. (2002). *Hierarchical linear models: Applications and data analysis methods*. Vol. 2nd ed. Thousand Oaks: Sage.

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

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica, 73*(2), 417-458.

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

Rothstein, J. (2010). Teacher quality in educational productionn: tracking, decay and student achievement. The Quarterly Journal of Economics, 175-214.

Rubin, Donald B., (2005) Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. Journal of the American Statistical Association, 100, pp. 322-331.

Rubin, D. B. (2008), For objective causal inference, design trumps analysis. The Annals of Applied Statistics 2: 808-840.

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., Saxon, A. M., Horn S. P. (1997) The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment in Millman, J. Grading teachers, grading schools: Is student achievement a valid evaluation measure? Thousand Oaks, CA: Corwin Press.

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

Scherrer, J. (2011) Measuring Teaching Using Value-Added Modeling : The Imperfect Panacea. NASSP Bulletin, 95(2) 122–140.

Schochet, P. Z. & Chiang, H. S. (2010). Error rates in measuring teacher and school performance based on student test score gains. Institute for Education Sciences.

Stipek, D. (2013). Using accountability to promote motivation, not undermine it *Ed Week* 33, 28-32.

Tekwe, C. D., Carter, R. L., Ma, C. X., Algina, J., Lucas, M. E. et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. Journal of Educational and Behavioral Statistics, 29(1), 11-36.

Todd P. E. & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. The Economic Journal, 113, f3-f33.

Walsh, E. & Isenberg, E. (2013). How does a value-added model compare to the Colorado Growth Model? MPR Working Paper: Princeton, NJ.

Wright, J. T., White, S. P., Sanders, W. L., & Rivers, J. C. (2010). SAS EVAAS Statistical Models. Cary, NC: The SAS Institute.

Wooldridge, J. M. (2009) Introductory Econometrics: A Modern Approach, 4th ed. Mason, OH: South-Western Cengage Learning.