Value-Added Modeling and Educational Accountability: Are We Answering the Real Questions?

Kimberlee C. Everson
Western Kentucky University

Value-added estimates of teacher or school quality are increasingly used for both high- and low-stakes accountability purposes, making understanding of their limitations critical. A review of the recent value-added literature suggests three concerns with the state of the research. First, the issues receiving the most research attention have not always been the concerns of greatest importance to theorists or critics. Second, there has been insufficient research on the interactions among various issues or assumption violations. Third, some of the big issues in value-added modeling have been challenging to address and may require educators to step back and answer some underlying philosophical questions about the nature of teacher and school quality.

Keywords: value-added modeling, educational accountability, teacher effect, school effect, teacher quality, teacher evaluation

A wave of legislation has been passed in states across the nation that requires student outcomes or growth to encompass at least a portion of a teacher’s evaluation—and often for high-stakes purposes (National Conference of State Legislatures, 2010). Baker et al. (2010) explained, “The Obama administration encourages states to make greater use of students’ test results to determine a teacher’s pay and job tenure” (p. 5). Policymakers and legislators in the United States, as well as in many other countries around the globe, have become increasingly enamored with school and teacher accountability based on student outcomes, and often for high-stakes purposes (e.g., Eckert & Dabrowski, 2010; Goldhaber & Hansen, 2008; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Organization for Economic Cooperation and Development, 2008; Sass, 2008).

Education researchers have acknowledged that evaluation of teachers or schools based on status (single time point) measures of student ability, such as end-of-year test scores, unfairly judges teachers and schools serving disadvantaged students (e.g., Organization for Economic Cooperation and Development, 2008). As an alternative, value-added (VA) modeling, a statistical methodology originating in economics and largely brought to the field of education by the work
of William Sanders (e.g., Ballou, Sanders, & Wright, 2004; Sanders, 1998; Sanders & Horn, 1994), has shifted the focus from end-of-year outcomes to student improvement. VA models estimate the relative effect of particular teacher and/or school assignments on end-of-year student test scores after either subtracting out or controlling for prior test scores. By explicitly accounting for differences in student prior ability by accounting for previous test scores, VA estimates of teacher or school quality are potentially fairer than status measures.

The goal of VA modeling is to compare a teacher’s or school’s student actual test score gains to the expected gains of the same student if he or she had not been assigned the particular teacher or school. These estimates (“effects”) are unbiased representations of a teacher’s or school’s contribution to student test achievement if the statistical model is correctly specified, statistical assumptions are met, and necessary measurement properties hold. Although researchers agree that these numerous conditions are never fully satisfied, there is no consensus on the degree of departure from these assumptions or the resulting impact on conclusions made about teachers or schools. If teachers and schools are going to be classified as high or low performing based on these measures, however, it is critical that we understand what current research exists and what future research may be necessary, so that inferences made about specific teachers based on the measures are as valid as possible.

Existing literature has referred to many concerns with VA modeling. In some cases, the purpose has been to discourage the use of VA modeling for accountability (VAM-A) purposes (e.g., Baker et al., 2010; Berliner, 2014; Hill, 2009; Martineau, 2010), although, in others, the tone has been more encouraging (e.g., Hanushek & Rivkin, 2010; Harris, 2009a; Harris & McCaffrey, 2010). Each author, however, has outlined only a subset of the research community’s concerns. A thorough analysis does not exist that encompasses all the major issues in the VA debate and the extent to which research addressing each issue has been explored. This analytical literature review sets aside the issue of whether student test scores ought to be used at all for school or teacher accountability and examines the existing methodological research related to VA modeling.

**Method**

*Criteria for Inclusion*

For this review, literature was selected using a disciplined selection protocol intended to focus on studies explicitly addressing VA modeling for educational accountability. Studies that addressed methodological concerns with VA modeling for the estimation of teacher or school effects for accountability were selected for this literature review. Both economic and statistical approaches were included, provided the focus was on model choice, statistical assumptions or interpretation, or measurement issues—as opposed to a primary concern with policy issues. To contain the scope of this review, literature was excluded that (a) addressed issues specific to postsecondary education, (b) compared VA estimates against other measures of quality unless the purpose was VA effect validation, (c) used measures other than test scores as outcome variables, or (d) focused on the potential effects or practicality of test-based evaluation. Non-U.S. studies were not excluded.
from the study if they met the inclusion criteria; however, the majority of the studies were conducted in the United States.

The literature that primarily addressed VA modeling for program evaluation, rather than VAM-A, was also eliminated. In addition, papers primarily addressing or comparing VA models with alternative growth model approaches such as the use of student growth percentiles (Betebenner, 2011) or principal axes (He & Tymms, 2014) were not included. Although it is true that these approaches are also sometimes referred to as “value-added models,” many statistical assumptions differ for these alternative approaches. Also excluded were papers addressing measurement and test construction issues that do not bear directly on VA estimates. Although these issues deserve further study and have relevance to VA modeling, this study’s purpose is to examine methodological issues that may have an impact on VA-based accountability in K-12 schools. VA models may, in theory, include nonlinear growth models, and one paper in the identified pool of literature used a nonlinear approach with promising results (Lopez-Martin, Kuosmanen, & Gaviria, 2014). However, because the remaining papers use a linear approach, this literature review primarily focuses on the issues addressing linear VA models.

Search Procedures

Education journals were searched using the ERIC database, and economics journals were searched using the EconLit database. To limit the review of the most recent literature, searches were limited to peer-reviewed journals, books, and economics working papers published from January 1, 2007, to October 1, 2015, that were indexed by October 1, 2015, reflecting the most research literature. Conferences, workshop proceedings, and reports were excluded. In ERIC, searches were conducted using the terms “school effect,” “teacher effect,” or “value-added.” In EconLit the same terms were used with the additional limitation that the VA articles must refer to “education,” “school,” or “teacher.” In addition, EconPapers was examined for any missed works that met the inclusion criteria. These two searches of economic papers included working papers that may have not met the peer review criteria in ERIC. The final literature pool of 99 works consisted of 1 book chapter, 71 peer-reviewed articles, and 27 working papers. Among these works, 55 were primarily written by economists, 37 were produced by education researchers or education statisticians, and 7 works had either dual education–economic authorship or a comparative focus.

Results and Discussion

Seventeen works were distinguishable from the rest of the pool in that they did not primarily present original research, but instead offered critical reviews or analyses of some of the important issues in VA modeling. Most of these works offered the authors’ reflections on or concerns with VA modeling (e.g., Amrein-Beardsley, 2008; Baker et al., 2010; Berliner, 2014; Gorard, 2011; Harris, 2009a; Hill, 2009; Martineau, 2010). A few, however, listed, but did not necessarily test, important assumptions required by VA models (e.g., Reardon & Raudenbush, 2009; Scherrr, 2011). All 17 works in this group had in common a purpose of listing the VA modeling problems, complications, issues, or underlying
assumptions that the authors suggested needed research. Together, these works describe the foremost current questions in VA modeling.

Table 1 summarizes the key issues identified by these papers. Each of these issues will be defined and discussed in depth in the remainder of this review. In some cases, an author was a contributor on multiple papers (e.g., Harris, 2009a, 2009b, 2010), but as the coverage was never identical, all articles by the author were included in the table. It is interesting to note from Table 1 that modeling decisions were apparently not of as great a concern to researchers as were statistical or measurement issues. A few of these papers included unique VA issues or statistical assumptions not addressed by any others (e.g., Harris & McCaffrey, 2010; Reardon & Raudenbush, 2009), but these issues will not be addressed in this review.

In this group of 17 papers, there was minimal disagreement as to either the state of the knowledge about VA or the facts related to specific statistical assumptions or methodologies. Whether each of these concerns was justified, or might lead to negative consequences, was generally not up for debate. However, there appeared to be disagreement as to the relative importance or potential impact of each issue. These authors’ opinions on specific issues will be reviewed within appropriate subsections within this review.

The most cited concerns with VA modeling in these 17 works included effect attribution ambiguities, instability of estimates, measurement error of test scores, the potential lack of an interval scale on test scores, sorting bias due to nonrandom assignments of students to schools or classes, and the validity of inferences made from the estimates (Table 1). As is shown in the remaining 82 works of literature in the selected pool, however, the recent research has not focused intensively on some of these areas of concern. For example, there has been minimal research that addresses the problem of the noninterval scale of scores and few addressing attribution ambiguities. In general, the research has addressed what can be studied and sometimes has neglected what should be studied. The mismatch between theorists’ questions and researchers’ efforts suggests many areas for increased research efforts in the future. The issues addressed in the remaining 82 works, the research-focused literature, will be discussed in detail. The next section addresses the VA research literature topically with modeling decisions first, followed by other statistical issues, and then measurement issues.

Modeling Decisions

Researchers who investigate VA modeling have a large number of decisions to make regarding the details of the statistical model used. A variety of questions about these modeling decisions were occasionally mentioned by theorists (Table 1). Most of the modeling issues addressed in the research literature pool centered on choices about the dependent and independent variables included in the model with the other questions addressed more minimally. In many cases, researchers did not investigate the appropriateness of various modeling choices or provided only brief justifications.

Within this pool of literature, researchers varied in their decisions regarding the appropriateness of including school effects in a model when the focus is on teacher accountability (e.g., Castellano, Rabe-Hesketh, & Skrondal, 2014; Ishii &
| Author                        | Modeling decisions | Statistical assumptions and interpretation | Measurement |
|-------------------------------|--------------------|---------------------------------------------|--------------|
|                              | Random vs. fixed effects | School vs. teacher effects | Pretest timing | Differential effects | Transparency | Imprecision and instability | Persistence | Attribution | Validity | Equal interval scale | Vertical scaling | Measurement error | Choice of testing instrument |
| Amrein-Beardsley (2008)       | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Baker et al. (2010)           | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Berliner (2014)               | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Gorard (2011)                 | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Hanushek and Rivkin (2010)    | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Harris (2009a)                | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Harris (2009b)                | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Harris (2010)                 | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Harris and McCaffrey (2010)   | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Hill (2009)                   | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Jennings and Concoran (2009)  | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Kelly and Downey (2010)       | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Linn (2008)                   | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Martineau (2010)              | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Misco (2008)                  | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Reardon and Raudenbush (2009) | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
| Scherrer (2011)               | X                  | X                                           | X             | X                       | X             | X                         | X             | X             | X         | X                   | X                 | X                   | X                     |
Rivkin, 2009; Papay, 2011; Rothstein, 2009), the necessity of modeling differential effects based on student traits (e.g., Condie, Lefgren, & Sims, 2014; Jakubowski, 2008; Lockwood & McCaffrey, 2009; Reardon & Raudenbush, 2009), and the use of fixed (e.g., Bacher-Hicks, Kane, & Staiger, 2014; Dieterle, Guarino, Reckase, & Wooldridge, 2015; Rothstein, 2007) or random effects (e.g., Briggs & Weeks, 2009; Lopez-Martin et al., 2014). Researchers disagreed regarding appropriate model complexity, with some disagreement as to whether preferring simpler, and therefore more transparent, models are preferable to models that are more complex (e.g., Gorard, 2008; Kelly & Downey, 2010). In some cases, effects were estimated from several models for comparison (e.g., Goldhaber, Goldschmidt, & Tseng, 2013; Lockwood, McCaffrey, Hamilton, 2007; Rothstein, 2009). Although it is possible that educational systems that use VAM-A make different choices than researchers make, the focus in the current article is on the modeling choices made for the purposes of researching VA methodology itself.

Table 2 reflects the modeling choices made, when specified, within the works chosen for this literature review. Although the most prevalently used model was a covariate-adjusted model with fixed teacher effects (e.g., Altonji & Mansfield, 2014; Stacy, 2014), reflecting the heavy representation of papers from economics, these choices were not universal. The discussion or use of each of these modeling choices will be examined.

| Decision             | Number of papers |
|----------------------|------------------|
| Form                 |                  |
| Gain models          | 6                |
| Covariate adjusted   | 47               |
| Both (comparison)    | 6                |
| Other                | 4                |
| Effect estimation    |                  |
| Fixed                | 32               |
| Random               | 22               |
| Both (comparison)    | 12               |
| Effect type          |                  |
| School only          | 18               |
| Teacher only         | 33               |
| School and teacher   | 10               |
| Teacher only plus    | 6                |

| Effect type          | Number of papers |
|----------------------|------------------|
| School only          | 18               |
| Teacher only         | 33               |
| School and teacher   | 10               |
| Teacher only plus    | 6                |

Gain and Covariate-Adjusted Models
The issue of modeling the prior test score using a gain or covariate adjustment was not an explicit concern of theorists (Table 1), but because it was investigated by some of the empirical literature in the pool, it is addressed here. Conceptually, the simplest VA models use gains (the difference between two test scores) as the
response variable (e.g., Guarino, Reckase, Stacy, & Wooldridge, 2014); other,
more complex models have used a multivariate approach, which allow multiple
outcomes, or multiple years of outcomes, to be modeled simultaneously (Briggs
& Weeks, 2011; Broatch & Lohr, 2012). Recent literature, however, has predomi-
nantly focused on covariate-adjusted models in which the response variable was
the current test score and a pretest or, more often, a prior year’s test score (or
several) was used as a covariate (Table 2). The more common use of covariate-
adjusted models has likely been a practical necessity because gains are less inter-
pretable when vertically scaled tests are not used (Briggs & Weeks, 2009).
Although most of the reviewed research used a covariate-adjusted model, a few
authors suggested reasons for using a gain model when a gain model is possible.
These included a belief in the gain model’s removal of the effects of bias due to
measurement error in the prior test scores (Ishii & Rivkin, 2009) and elimination
of the correlation of the lagged covariate (previous test score) with the error term
(McCaffrey, Sass, Lockwood, & Mihaly, 2009).
A simulation study by Zamarro, Engberg, Saavedra, and Steele (2014) indi-
cated that covariate-adjusted models produced more accurate rankings of teachers
than did gain models, but they were not always better at recovering the simulated
student achievement to teacher VA correlation. Schochet and Chiang (2013),
however, found that, so long as a vertically scaled test was used, gain and covari-
ate-adjusted approaches resulted in similar estimates. Similarly, Sass, Semykina,
and Harris (2014) found strong correlations between the effects estimated from
gain and covariate-adjusted models. However, they noted that even small differ-
ences can be practically important. Although the lack of a set of comparison
effects from a randomized trial makes the calculation of actual bias impossible,
these authors assumed that the smallest total bias is likely to occur in models
requiring the fewest assumptions, leading them to favor the gains model.
Several authors (Ishii & Rivkin, 2009; Kane & Staiger, 2008; Koedel & Betts,
2011; McCaffrey et al., 2009) conducted analyses using both gain and covariate-
adjustment models, but the purpose of their studies was not to compare estimates
from these models to each other. Instead, they investigated other VA issues within
each of these models. Lockwood, McCaffrey, Hamilton, et al. (2007) did compare
estimates from gain and covariate-adjustment models and found the estimates
were correlated \( (r > .60) \). The purpose of their analyses, however, was not to make
conclusions about gain versus covariate-adjusted models but to show how the
issue of model choice compares with the issue of testing instrument choice. At
this point, covariate-adjusted models are most frequently used in the literature,
likely due to their flexibility in use when tests are not vertically scaled rather than
because of empirical or theoretical reasons suggesting their use is preferable.

**Random and Fixed Effects**

Theoretical and statistical reasons for making decisions between fixed and ran-
dom effects in statistical models are well-documented in statistical textbooks and
literature. Generally, fixed effects are used when inference is only intended to the
groups (e.g., classrooms, schools) included in the data collection, but random
effects are used when inference to a larger population is intended (Snijders, 2005).
Other reasons for using a random effects approach may include the desire to add

41
group-level covariates to the model or the existence of small within-group sample sizes. However, additional assumptions are necessary when random effects are estimated, and a larger number of groups is required (Timmermans, Snijders, & Bosker, 2012).

The first issue examined by researchers was whether the two sets of effects are practically and statistically different, regardless of assumptions that may have been violated. Deming (2014) conducted a small study in which he found effects estimated using random and fixed approaches were similar. Jakubowski (2008), however, compared fixed and random school effects and found that they were noticeably different, although they became more comparable when small classes were removed from the data set.

In comparing random and fixed effects, most authors, however, were concerned not just with whether the two sets of effects were the same but with whether greater bias in effect estimates resulted from one set of effects than the other. In the included VA literature, economists usually used fixed effects, and education researchers often used hierarchical models that include random, or shrunken, effects (see Harris & McCaffrey, 2010). Clarke, Crawford, Steele, and Vignoles (2010) compared the assumptions, advantages, and disadvantages of random and fixed effects models for estimation of school effects. The authors suggested that economists prefer fixed effects models, partly due to concerns about the additional assumption required by random effects models that the random intercept (teacher or school effect) be uncorrelated with other independent variables. The authors explained, however, that random effects models, when justified, have several advantages: (a) an ability to model differential effects through the use of random coefficients, (b) greater flexibility in the research questions that may be addressed, (c) improved efficiency of the estimates (narrower confidence intervals), and (d) shrunken estimates, which reduce the impact of “rogue” effects due to small group sizes or a large within-group to between-group variance ratio (also see Ray, McCormack, & Evans, 2009).

Although the biggest reason for not using a random effects model is the strong potential for violation of the random effects assumption, the authors suggested that bias due to violation of the regression assumption is a problem for both random and fixed effects models. They indicated that this problem generally overshadows bias introduced by violation of the random effects assumption, eliminating the advantages of using a fixed effects model. The authors concluded that random effects models are justifiable if the data are rich (contain a large selection of covariates) and there is a good “theoretical understanding of school selection mechanisms” (Clarke et al., 2010, p. 14). Thus, the more accurately we can model the factors that affect selection to schools, the more advantages there are with using random school effects, in spite of violations of the random effects assumption.

Some of the remaining literature supported Clarke et al.’s (2010) conclusion that random effects models and fixed effects models each may be preferable in different situations. Several authors, for example, found evidence of greater bias in random effects when there is sorting bias. For example, Castellano et al. (2014) found that when models control for school socioeconomic status (SES) or other factors that also predict the quality of teachers assigned to the school, the random
effects assumption is violated, and this can lead to bias (see also discussion under “Covariate choice”). Similarly, Guarino, Maxfield, Reckase, Thompson, and Wooldridge (2015) found that random effects resulted in more bias than do fixed effects when nonrandom assignment of students to teachers exists. One workaround for the potential violation of the random effects assumption was attempted by Manzi, Martin, and Van Bellegem (2010). The authors agreed that the random effects assumption is likely to be violated in VA models and attempted to adjust for the problem adding parental education as an instrumental variable. In another study, random effects appeared better in a simulation study when sorting to classrooms did not occur, but fixed effects were preferred in the presence of sorting (Zamarro et al., 2014).

Other studies, however, indicated less bias when a random effects approach was used. In an examination of the VA estimates of teachers whose student achievement is harder to estimate due to small sample size or low SES and low prior ability, random (shrunken) effects were found to be more precise, and unsurprisingly smaller in absolute value. However, the use of random effects did not change the probability of a teacher being considered high or low performing (Herrmann, Walsh, Isenberg, & Resch, 2013). Using a series of simulations, Henry and Rose (2015) found that random effects–based models consistently resulted in less bias than fixed effects approaches when various model assumptions were violated. In addition, a random effects approach falsely identified a smaller percentage of poor performing teachers whether or not model assumptions were violated. However, as found by Herrmann et al. (2013), results were inconclusive as to whether fixed or random approaches were preferable when the question of interest was the percentage of teachers who would shift from lowest to highest quintile, or vice versa, over 2 years. Karl, Yang, and Lohr (2013) found a correlated random effects model can be useful for investigating the impact of missingness on VA teacher rankings. In another study, Guarino, Reckase, and Wooldridge (2012) found that no estimation approach is better in all situations. Clarke et al. (2010) implied that the decision should be based on research purpose, desired interpretation, and the probable justification for the random effects assumption in a given model.

Overall, this review of the literature suggests that the choice of random versus fixed effects may most appropriately depend on the degree of sorting bias that is present, with fixed effects resulting in less bias when sorting is more prevalent and random effects more precise when there are many covariates or smaller class sizes. The choice of effect type, however, does not appear to affect categorization of teachers to high- or low-performing status.

Covariate Choice

VA models always include prior achievement or ability either as a covariate in the model or by subtracting it from final achievement. In addition, VA models sometimes include additional covariates at the student, family, classroom, or school level. These covariates “control for” (adjust) school or teacher estimates of effectiveness to account for the differences in school or classroom compositions, purportedly making effect estimates less biased. One commonly addressed issue in the literature has been which, or how many, covariates to include in the model.
Many researchers have suggested that using a large number of covariates reduces the risk of effect bias due to attributing the effect of a missing covariate to a teacher or school (Clarke et al., 2010; Dearden, Miranda, & Rabe-Hesketh, 2011; Levine & Painter, 2008). Dearden et al. (2011), for example, found that failing to include mother’s education in the model, after inclusion of a rich set of student demographic, academic, and socioeconomic variables, resulted in biased estimates of school VA. The authors concluded that factors outside the school, which are frequently not included in VA analyses, can bias estimates. Johnson, Lipscomb, and Gill (2013) found that VA estimates were highly correlated regardless of covariate inclusion decisions, but that categorizations of teachers (e.g., into quintiles) was often highly sensitive to covariate specification. The sensitivity to inclusion/exclusion of a specific covariate varied depending on the distribution of student traits within a district.

Others, however, have intentionally limited the use of covariates for practical or theoretical reasons. Rothstein (2009), for example, suggested that rarely is more than one previous test score added to the model due to data limitations, although he finds that bias is reduced by using several years of prior scores. Lenkeit (2013) suggested even prior achievement is not always easy to obtain. She compared models that used only student covariates (contextualized attainment models) to those that also included prior achievement, and confirmed that prior achievement is the most important factor in predicting school achievement. Although Harris and McCaffrey (2010) agreed that adding covariates is challenging, they addressed a more fundamental concern. The authors argued that including student-level covariates may reinforce differential expectations for students of different demographic groups, suggesting that adjusting estimates of teacher or school effectiveness to account for student types may not always be appropriate—all should be held to the same standard (see also Schochet & Chiang, 2013).

Gorard (2008) suggested that another cost of adding background covariates to a model is that we can no longer use the estimates to tell whether schools or teachers with a higher or lower proportion of students with that trait are equally effective. Another issue may be whether a teacher or school should adjust for or be held accountable for challenges within the school, such as a high truancy rate (Kelly & Downey, 2010). Similarly, adjusting for SES at the between-school level may unfairly reduce an estimate of school VA if stronger teachers are drawn to higher SES schools, thereby making too big an adjustment for SES (Castellano et al., 2014). Essentially, these authors suggested that adding covariates may inappropriately alter the comparability of effect estimates of various subgroups of schools or teachers.

Castellano et al. (2014) indicated that inclusion of group-mean student covariates led to bias when between-school effects rather than within-school effects were of interest. They indicated that econometricians have attributed these differences to a confounding of the group-mean covariate with unobserved traits (“Level 2 endogeneity”), in contrast to education researchers who have attributed the problem to the collective impact of peers or other factors within schools (“contextual effects”). Using a Housman–Taylor estimator that first estimates the coefficient for the covariate using fixed effects approach, and then correctly adjusts for SES in future models, they solved the problem of
bias due to Level 2 endogeneity, but did not solve problems of bias due to contextual effects. By contrast, the authors found that accounting only for contextual effects by including school or class means as covariates in addition to student-level covariates, as is often done by education researchers, can lead to Level 2 endogeneity bias. The authors favored the Housman–Taylor approach and concluded it produces school effects (or teacher effects) that estimate contributions of school practice and school context (or teacher practice and classroom context) jointly. They indicated that estimating effects that only include the impact of schools or teachers remains an unsolved problem, but they suggested that the Housman–Taylor approach or mean-covariate approach might produce good enough estimates if one source of bias (Level 2 endogeneity or contextual effects) is relatively small.

Another common finding was that adding covariates beyond prior test scores did not make a practical difference in effect estimation (e.g., Harris & McCaffrey, 2010; Levine & Painter, 2008; Lockwood, McCaffrey, Hamilton, et al., 2007; Schochet & Chiang, 2013), but this result was not universal. Ferrão and Goldstein (2009) found that adding covariates representing SES and prior achievement to a null model significantly changed the estimates, but adding additional covariates (e.g., gender, special education needs, whether grade was repeated) beyond that made little difference. Papay (2011) examined a set of models with a variety of specifications with respect to individual, classroom, and school covariates and found that covariate selection decisions made some difference in ranks of effect estimates ($r_s > .77$) but not as much difference as was found by varying the assessment used ($.52 < r_s < .65$).

Jackson (2012) found that accounting for academic tracking of high school students was important after accounting for prior achievement, parental education, ethnicity, and gender. Similarly, Protik, Walsh, Resch, Isenberg, and Kopa (2013) found that track indicators or classroom characteristics reduced bias after controlling for prior achievement, special education status, and various socioeconomic and demographic student traits. Jakubowski (2008) found correlations in school effect estimates using only prior achievement, gender, and dyslexia status with models also including SES ranged from .74 to .82. Although this range is high, it may mean that some individual teacher evaluations could depend on the covariates included. Lefgren and Sims (2012) found that multisubject weighted VA models meant to predict elementary teachers’ reading VA predicted future reading-only teacher VA, especially when only 1 or 2 years of data were available, suggesting same subject-only prior test scores may not be sufficient to estimate teacher quality.

Kersting, Chen, and Stigler (2013) found their estimates were particularly insensitive to covariate inclusion decision, with single prior test score models producing similar results to other specifications, which included two prior test scores and/or student covariates. The authors theorized that this insensitivity may have been due to unique characteristics of their sample. Student test scores were highly correlated both across years and with student covariates. Thus, there was little extra information provided by including more than one prior test score. They suggested that these strong relationships may have been due to high reliability in the test scores and warned that other districts may not get similar results.
Timmermans, Doolaard, and de Wolf (2011) categorized the different approaches to adding covariates when school effects are of interest. These categories included models with no controls, those with only prior achievement as a control, those that also included student-level covariates other than achievement, models that added to those controls nonmalleable school-level compositional covariates, and models that additionally included other school-level covariates. The authors found correlations in school effects across these modeling choices that ranged from .630 to .959, although the authors suggested the high correlations might be partially a result of shrinkage toward the mean due to small class sizes. In spite of the high correlations, the authors found that individual schools often shifted from being under-, average, overperforming depending on the modeling choice.

The inclusion or exclusion of specific covariates may depend on first answering some philosophical questions. As suggested by Kelly and Downey (2010), we are not always sure with whom a teacher or school’s students should be compared. Without first answering this question, decisions may continue to be made inconsistently across researchers and risk being based more on data availability than on purpose.

School and Teacher Effects

As indicated by Table 2, some VA models estimated only school effects, others estimated only teacher effects, and others included both. In most cases, the choice depended on purpose. For school accountability, only the school effect needed to be modeled. For teacher accountability, however, the choice was less clear. When school effects and teacher effects were both included in the model, or when group-mean centering was used in a random effects approach, a teacher’s estimate of effectiveness became relative only to other teachers in the same school because the school mean was accounted for (e.g., Castellano et al., 2014; Ishii & Rivkin, 2009; Papay, 2011; Rothstein, 2009). Others, however, found that limiting comparisons to within schools may not be appropriate for accountability as much of the school-level effect is attributable to teachers (e.g., Hanushek & Rivkin, 2010; Jackson, 2009).

This decision to include or exclude school effects, moreover, can have a large impact on teacher effects (Papay, 2011). In addition, teachers who switch schools can complicate estimation models that use multiple years. Goldhaber and Hansen (2013) specifically excluded school effects because they wanted to look at individual teacher effect change over time, and teachers often switched schools. By contrast, two studies demonstrated that, by capitalizing on school changers in their model, they were able to estimate teacher effects at the across-school level while still estimating school effects (Bacher-Hicks et al., 2014; Chetty, Friedman, & Rockoff, 2014a). At present, however, there is no clear solution to the problem of estimating teacher and school effects simultaneously without reducing teacher effects to within-school comparisons, and most models interested in teacher effects did not also include a school effect (Table 2).

Pretest Timing

Another VA design decision is whether to use as the measure of incoming ability a pretest taken by the students in the fall or the previous end-of-year testing. Within the selected literature, only a few studies modeled twice annual testing
Value-Added and Educational Accountability

(fall and spring) instead of annual testing (Kyriakides & Creemers, 2008; Palardy, 2010). Palardy (2010) demonstrated that student growth, and thus teacher effects, in the following school year were affected by whether or not the summer loss was included in the model.

However, Baker et al. (2010) argued that there are practical problems with measuring students in the fall. Testing students twice a year takes up time and tempts teachers to "game the value-added measures" by discouraging the students from doing well on the fall test, "if only by not making the same extraordinary efforts to boost scores in the fall that they make in the spring" (p. 15). In addition, the authors asserted that fall to spring testing requires tests to be spaced as far apart as possible, which means that spring tests may happen too late for the policy purpose of terminating contracts for the next year. Future research may confirm both the potential extent of the practical concerns voiced by Baker et al. (2010) and the statistical issues demonstrated by Palardy (2010).

Modeling Differential Teacher Effects

Another concern related to VA models is that teachers or schools may not be equally effective with all types of students (Everson, Feinauer, & Sudweeks, 2013; Loeb, Soland, & Fox, 2014). If they are not, then some researchers wonder whether the model should reflect this disparity (Condie et al., 2014; Stacy, 2014). Jakubowski (2008) examined differential school effects by adding a random slope to his random effects model. He found that effects from the models with and without the random slope have correlations between .84 and .98. Lockwood and McCaffrey (2009) conducted an empirical study and found that, in most cases, allowing for heterogeneity had little impact on teacher effects, despite theoretical justification for believing that teachers teach different students better than others. They suggested, however, that more research should be done as some teacher estimates in their study were altered significantly when heterogeneity was ignored. Loeb et al. (2014) conducted a study of the stability of teacher VA across student subgroups (English learner [EL] vs. non-English learner) and found that most teachers who are effective with one group were effective with the other. However, some teachers were more effective with ELs than with non-ELs.

Reardon and Raudenbush (2009), however, used the results of simulations to suggest that any existing heterogeneity should be modeled unless there is strong reason not to. In addition, they found that ignoring this heterogeneity was more problematic in some situations than in others—it interacted with the level of student sorting and the functional form of the model. Condie et al. (2014) found that teacher effects are clearly heterogeneous across subjects and student types, leading to misrankings of teachers 15% to 25% of the time when homogeneity is assumed as is done in a typical VA model. They find that even with the misrankings, a policy that fires the bottom 10% of teachers would improve student achievement. However, by acknowledging heterogeneity and correctly matching teachers to student types instead, achievement could be improved more dramatically than by firing teachers using a homogeneous VA effect model. Chetty, Friedman, and Rockoff (2014b) also found that teacher effects were heterogeneous and that the long-term effect of a teacher, in particular, was dependent on student SES. Another study focused on elementary
teacher quality and found evidence of effect heterogeneity across subjects (Goldhaber, Cowan, & Walch, 2013).

Stacy (2014) suggested examining VA variance for each teacher as a measure of heterogeneity. He found evidence that adding information about variance to a teacher ranking system produces teacher rankings highly correlated with mean VA-only models (rs > .90). In addition, he found that student, classroom, and teacher traits explained little of the variation in the variance in teacher VA. Because adding VA variance is a simple process, he suggested more research on this potential approach.

These papers provided some empirical evidence that modeling teacher effects as nonconstant across student types can affect estimates and provide greater information about teachers. However, effects that are modeled as heterogeneous can be challenging to interpret; thus, future research may be helpful in determining in which cases modeling heterogeneity, if it exists, is most critical.

**Transparency**

Lack of transparency can refer to the unavailability of information about modeling details to outside parties. For example, the Education Value-Added Assessment System has not made specific details of the methodology available for review (Amrein-Beardsley, 2008). However, more often, these modeling details are available but not necessarily easily understood by education professionals. There is a tension in modeling decision making between model transparency, or simplicity, so that teachers, administrators, and policymakers might understand what a teacher or school effect represents, and complexity in order to obtain the best estimates (e.g., Koedel & Betts, 2011). Kelly and Downey (2010) argued that a model must be simplified to the point that “outside expertise” (p. 193) was not required to interpret it, even at the cost of losing any improvements to the model that come with greater complexity. Gorard (2008) suggested, “The solution to all of these issues is not a more complex value-added analysis. The solution lies in re-thinking what it is that we want value-added analysis to achieve” (p. 183).

Others implied, however, that estimates must be as precise and accurate as possible, regardless of complexity. For example, Timmermans et al. (2012) found that modeling student mobility affects estimates and can be helpful, in spite of added complexity. Part of the reason for this disparity in belief is that the models have various purposes to various stakeholders. “And despite their complexity, the measures fail to respond adequately to competing legitimate demands: from the public for interpretability; from teachers for usefulness; and from policy-makers for accountability” (Kelly & Downey, 2010, p. 196). The level of transparency versus complexity may depend on this purpose, and it may not always be clearly defined.

**Summary**

The literature selected for this review reflected diversity in the modeling choices that are made, reflecting the complexity and large range of possibilities in creating VA models. Although there was a set of literature addressing the advantages and disadvantages of many of these choices, many of the papers did not
provide justification for modeling decisions. In addition, few papers examined the interaction of many of these modeling choices. One exception not addressed in prior subsections is a paper by Guarinio et al. (2012) that used simulations to compare a variety of methods, including random and fixed effects and gain and lagged models, under a variety of conditions, and found that no one method was superior to the others in all situations, although, in many cases, a dynamic least squares method performed best. The authors suggest that there is “much left to investigate,” including the interaction between modeling choices and statistical and measurement issues (Garinio et al., 2012, p. 32).

**Statistical Assumptions and Interpretation**

This section addresses the most commonly expressed concerns with statistical assumptions and interpretation in VA modeling. Although there are other assumptions and statistical problems that have been touched on either within the reviewed literature or by others in the past, included here are only issues explored by at least two works within the literature pool.

**Sorting Bias**

The nonrandom assignment of students to schools and, within schools, to classrooms or teachers, was the most widely acknowledged potential limitation to causal interpretation of teacher or school effects (Table 1). Baker et al. (2010) stated,

> In sum, teachers’ value-added effects can be compared only where teachers have the same mix of struggling and successful students, something that almost never occurs, or when statistical measures of effectiveness fully adjust for the differing mix of students, something that is exceedingly hard to do. (p. 11)

A teacher whose assigned students are ripe to make big gains during the year will be at an advantage, unless those propensities can be accounted for in the model. In the reviewed articles, authors attempted both to determine to what extent sorting (at the school or classroom level) occurs and to understand to what extent that sorting affects VA teacher or school effects.

**Classroom or teacher sorting.** Within this literature pool, inquiry into classroom sorting began with Rothstein’s (2007) working paper, subsequently published in 2010, in which he developed a falsification test for the assumption that students are randomly assigned to teachers. He found that a student’s fifth-grade teacher assignment predicted fourth-grade outcomes, which should only occur if students are nonrandomly assigned to fifth-grade teachers based on their fourth-grade outcomes. According to Rothstein, if classroom assignment is based on unobserved time-varying factors, then VA estimates of teacher effects are biased and cannot be assumed to be causal. Much of the recent literature addressing classroom-level sorting was in reaction to this Rothstein study.

Koedel and Betts (2011) replicated Rothstein’s (2007, 2010) analysis on a different data set and confirmed his findings. The authors then extended the analysis using more complex models and multiple years of data. In all cases, although
evidence of sorting bias was found, this evidence was smaller than when replicating Rothstein exactly. Using 3 years of data for teachers almost completely eliminated the effects of sorting bias. The authors suggested, however, that classroom sorting could potentially be used to manipulate VA measures in high-stakes settings. Altonji and Mansfield (2014) found that, with large samples, teacher VA should not be biased, even if unobservable confounders exist, so long as measures of average student traits and classroom characteristics are included in the model. They suggested that this result partly explains the promising findings of Koedel and Betts.

Several authors have tested Rothstein’s (2010) falsification test as a means of identifying models that are likely biased due to sorting. Sass et al. (2014), for example, found future teacher assignments had an effect on student outcomes, using a variety of VA models, suggesting teacher assignment is, to some degree, related to student ability. Kinsler (2012a) found, however, that the Rothstein falsification test does not work well with small samples or when the model is misspecified. Goldhaber and Chaplin (2012) found that Rothstein’s results could be due to other reasons than nonrandom assignment, such as nonlinearity in the relationship between the lagged and double-lagged test scores rather than due to teacher sorting bias. Finally, Guarino et al. (2014) demonstrated that Rothstein’s (2007, 2010) falsification test has a high risk of identifying models as resulting in bias when they do not in fact produce bias. This concern appears to especially be a problem when student tracking is used to form classrooms, but those classrooms are randomly assigned to teachers.

Also building on Rothstein’s (2007, 2010) falsification study is Rothstein’s (2009) article in which he reasoned that the more unmodeled information the principal has and uses for classroom assignment, the larger would be the bias in estimation of teacher effects. He found that, even if the principal knew little beyond what was observed and included in the model, biases were larger than desired. Similar to Koedel and Betts (2011), however, Rothstein found that this bias was minimized by increasing the number of prior year test scores used as covariates. Nonetheless, he concluded that even the best models contained bias. He theorized that in high-stakes settings, teacher pressure might tempt the principal to weight student potential more heavily based in sorting, thereby increasing bias.

In a more direct attempt to study the effect of nonrandom assignment on VA estimates of teacher effectiveness, Kane and Staiger (2008) designed a randomized experiment in which 78 pairs of teachers in Grades 2 to 5 were randomly assigned between two classrooms. The authors found previous teacher VA estimates predicted classroom gains in the random-assignment year, providing validity evidence favoring interpreting VA estimates as indicators of some stable aspect of teacher quality. Rothstein (2009), however, critiqued Kane and Staiger for several reasons, including low power and the unusual nature of the sample.

Examining potential sorting bias from another perspective, Chetty et al. (2014a) demonstrated that teacher shifts across schools resulted in changes in mean student test scores at the prior school that were predicted by the loss of the teacher. Schools losing a high VA teacher, for example, had lower student test scores in that subject the following year, suggesting that inferences about teacher contributions based on teacher VA effects have some validity, in spite of sorting
bias. There has been some debate between the authors and Moshe Adler (2014), however, regarding whether the study correctly accounted for the intertemporal instability of VA effects and other methods (Chetty, Friedman, & Rockoff, 2014c). Bacher-Hicks et al. (2014) replicated the Chetty et al. (2014a) study using data from Los Angeles and came to the same conclusion that teacher VA estimates are good predictors of achievement. Although not identified using the search criteria employed in this literature review, an unpublished working paper (Rothstein, 2015) also replicated Chetty et al. (2014a) using a different sample.

Rothstein (2015) raised concerns with the approach’s methodology, suggesting that sorting bias was not eliminated. He indicated, however, that problems resulting from using VA scores for personnel decisions may not be severe. Furthermore, Bacher-Hicks et al. (2014) found Rothstein’s (2015) methodological concerns to be unwarranted. Another quasi-experiment (Deming, 2014) capitalized on students accepted or denied acceptance to schools out of their assignment area based on a lottery system and found no evidence of sorting bias in VA estimates, although the author acknowledged the sample was small and represented only students who wanted to shift schools.

Reardon and Raudenbush (2009) found simulation-based evidence that violations of the random assignment assumption did not significantly impact effect estimates unless those effects were heterogeneous. On the other hand, Dieterle et al. (2015) found evidence that sorting to classrooms within a school is significant and affects estimates. They found, however, that the degree of sorting varies in degree from school to school.

In summary, most of the reviewed literature appears to suggest that in many cases differences in classrooms due to sorting do not appear to introduce significant bias in effect estimates, especially when several years of data are included in the analyses. As addressed in the section examining random and fixed effects, it may be that the real question is which model to use when classroom sorting is a significant risk.

School sorting. The issues surrounding student sorting to schools are in some ways different from those surrounding classroom-level sorting as the student assignment mechanism works differently at the school than at the teacher level. The recent research evidence addressing the impact of school-level sorting, however, was sporadic and limited.

Gorard (2008) found that VA models did not remove the bias due to differential student ability across elementary schools in England. No low-attaining schools were found to have average or high VA estimates, and no high-attaining schools were discovered to have low VA. The author concluded that this relationship is too strong to occur just due to chance and reflects school sorting bias in VA estimates. Jackson (2009) used desegregation data to find evidence that better teachers tended to end up in advantaged schools, but did not convincingly separate that effect from bias in the teacher effects themselves due to differential student ability in lower performing schools. Similarly, Ready (2013) found that VA models did not remove the correlation between achievement and growth at the school level, although the form of the relationship varied across a variety of analytic techniques he employed.
Everson

Van de Grift (2009) suggested that VA measures in his data were only reliable for schools with higher achievement as there was so much missing data in schools with lower achievement. Levine and Painter (2008) presented the most optimistic viewpoint on school sorting by finding catch-up effects when students changed schools, suggesting that school effects did represent school quality. Similarly, Altonji and Mansfield (2014) were optimistic that bias at the school level with large enough samples is not a problem, so long as appropriate covariates that contain averages of student traits are included. The innovative methodological approaches used by each of these six papers suggest one of the challenges for research related to school-sorting bias—framing the problem is not straightforward.

Imprecision and Instability

Stable estimates remain constant across time, testing instrument, model form, or other factor. Precise estimates have small standard errors. Baker et al. (2010) explained that these related properties result in uncertainty as to whether the truly weakest teachers were the ones identified as weakest. The authors suggested that studies that examined instability “suggest that there is not a stable construct measured by value-added measures that can readily be called ‘teacher effectiveness’” (Baker et al., 2010, p. 13). It might be more precise, however, to say that VA estimates may not yet effectively measure a stable trait of teacher effectiveness, if there is one. The reviewed literature examined effect stability from the several dimensions, including time and instrument choice.

Using a variety of models, McCaffrey et al. (2009) estimated single-year teacher effects for consecutive years and found low correlations between years. Only a third of the teachers who were ranked in the lowest quintile remained there the next year, and 10% of those low-ranking teachers moved to the top quintile. The authors found that averaged teacher effect estimates across several years were more stable. McCaffrey et al. argued that similar instability in other professions provides evidence that teacher quality truly changes over time and does not suggest a problem with the estimates. Goldhaber and Hansen (2013) used a time series approach to demonstrate that, although there is a component of the teacher effect that is stable across time, that component is small. Ferrão (2012) ranked teachers by quartile using a VA model and found that only about 65% of teachers stayed within the same level of performance (quartile) for at least 2 years and very few remained within the same quartile for all 3 years of the study. Kersting et al. (2013) found only 30% changed among three performance groups across years, and only 1% shifted between the lowest and highest groups. Their approach, however, used statistical significance to sort teachers into categories rather than the more commonly used quartile or quintile approach. They also noted that student test score correlations across cohorts were high, suggesting that teachers were assigned similar students from year to year in their sample. These authors found that sample size, however, was a key predictor of stability and suggest more research is necessary to set guidelines for minimum sample sizes when VA estimates are used.

Another important observation made by both McCaffrey et al. (2009) and Stacy, Guarino, Reckase, and Wooldridge (2013) was that bias and stability may have an inverse relationship. When there is persistent sorting, as when a teacher
specializes in a specific type of student, then that teacher’s effect may be quite stable. It will, however, be biased due to the sorting. On the other hand, Berliner (2014) suggested instability in effect estimates might be partially caused by unaccounted for sorting (exogeneity). Sass (2008) presented a briefer version of the results from McCaffrey et al. (2009) with some additional policy insights. Goldhaber and Hansen (2013) reflected that even intentional principal favoritism that consistently rewards certain teachers with preferred students can increase stability.

Similar evidence of at least some intertemporal instability was consistent across the reviewed literature (e.g., Goldhaber & Hansen, 2008; Gorard, 2011; Kyriakides & Creemers, 2008; Leckie, 2009). Goldhaber and Hansen (2008) suggested that there was slightly more stability in their study than in others because they used multiple years of data. Ishii and Rivkin (2009), however, indicated that averaging teacher effects over several years may result in a “trade-off between obtaining more precise estimates and recognizing that teacher quality is not a fixed characteristic” (p. 535). Thus, these authors suggest that teacher quality should not be stable—teachers should try to improve, and multiple year measures may mask that improvement. Schochet and Chiang (2013) focused on the problem from the perspective of effect precision and found 26% of teachers were erroneously identified as differing from the mean in a simulation study using 3 years of data.

VA estimates have also been found to be unstable across testing instrument (Lockwood, McCaffrey, Hamilton, et al., 2007; Papay, 2011), subject (Goldhaber et al., 2013), and statistical model (Newton et al., 2010). In addition, Leckie (2009) found that schools were ranked differently depending on whether the students’ earlier schools had been taken into account. Sanders, Wright, and Langevin (2008) found that teacher VA showed some instability for teachers who made shifts to different schools, but that this instability decreased the second year in the new school, suggesting that VA estimates largely reflect a stable teacher trait.

The reviewed research consistently found at least some effect instability, especially when comparing teacher effects across years, and in general, the literature addressing instability is most often negative toward the potential of VA modeling. However, the impact of other sources of instability, and the interaction of effect instability with other VA issues, has not been adequately addressed in the research. In addition, it is unclear to what degree instability in teacher effect estimates is due to real differences in the trait being measured over time, other factors, or a problem with the estimates themselves.

**Persistence**

The concept of persistence, or long-term effect, was examined in two slightly different ways in this literature pool. First, VA models can be built to reflect an assumption either of complete or of variable persistence of the impact of prior inputs on a given year’s student outcomes. When models were estimated both ways, conclusions about teachers were found to vary to a degree that may be practically important. Although differences in estimates were not generally statistically significant, different teachers were often ranked below or above average depending on decisions about modeling persistence (Lockwood, McCaffrey,
Mariano, & Setodji, 2007; Sass et al., 2014). When school instead of teacher effects were estimated, Briggs and Weeks (2011) found that, although the effect of ignoring incomplete persistence was smaller for schools than for teachers, there was little justification for using a complete persistence model.

In several papers in this literature pool, however, persistence referred specifically to the ability of high or low VA estimates themselves to predict future ability in the same subject, or possibly on other outcomes. The focus was not on modeling persistence or lack thereof but on empirical questions about whether or not teacher or school impacts persist, or continue to have impact, in future years. If the impact does not persist, then VA estimates are probably only estimating short-term inputs, such as teaching to the test. Most of the reviewed literature was consistent in suggesting that VA teacher effects only partially persist; they fade out over the first couple years quite quickly. For example, Jacob, Lefgren, and Sims (2008) found that teacher effect on mathematics and reading scores quickly diminish with only one fifth of the effect remaining after 1 year and about one eighth after 2 years. Kane and Staiger (2008) estimated a more modest effect erosion of about 50% per year for the 2 subsequent years of instruction. Kinsler (2012b) found that only about a third of the effect survived a year.

In contrast to the more negatively focused studies regarding the persistence of teacher effects, one recent study found that teacher VA predicted long-term outcomes. Here, it was found that elementary teacher VA predicted class-level means in college attendance, teenage pregnancy, and future salary (Chetty et al., 2014b). Research that attempts to replicate this single study may be helpful in evaluating this promising result, particularly as it suggests that a teacher may have long-term impact on tangentially related outcomes, in spite of the other studies suggesting the teacher’s impact erodes quickly on same subject outcomes.

Attribution

Even if an estimate of teacher or school quality is unbiased and genuinely reflects learning gains by students within that classroom or school, there is the possibility of attribution problems. The fact that a class of students makes unusual gains does not mean that the teacher was the cause of those gains, even if background variables are controlled for. Much of the literature referring to problems of attribution, however, was theoretical or conjecture.

Baker et al. (2010), for example, hypothesized that some of the possible influences that are confounded with teacher inputs include outside tutors, class sizes, counselors, and specialist support. Ishii and Rivkin (2009) suggested that parental response to perceived teacher quality (e.g., by providing extra tutoring if the teacher is of poor quality) is a particularly problematic source of potential attribution problems. These authors theorized that even random assignment cannot compensate for parental response to teacher quality, and it is difficult to measure. Rothstein (2009) also acknowledged the attribution problem and suggested using the term “classroom effect” instead of “teacher effect” to prevent misunderstanding (p. 539).

A few authors, however, addressed the problem of attribution empirically. For example, Koedel (2009) demonstrated one potential approach to confirming the existence of confounding influences on student gains by showing that student
reading gains can be attributed to both reading and math teachers. Hock and Isenberg (2012) were innovative in presenting two approaches for separating inputs by teacher when teachers have joint responsibility for a group of students. Isenberg and Walsh (2013) built on this work and suggested a model for the particular attribution problem of coteaching, the Full Roster-Plus Method, which gives both teachers equal credit for all students, counting students twice for teacher effect estimation but only once when examining student traits. In addition, experimental or quasi-experimental studies such as Kane and Staiger (2008) and Deming (2014) suggested that, at the least, there is evidence that the differences between classes are causally related to things happening at the individual classroom level and not due to school factors.

**Validity**

In a piece included within the theoretical papers (Table 1), Hill (2009) suggested that inferences from VA scores should be held to the same standard of validity as that used for test scores. In particular, VA scores should reflect teacher quality, should be accurate, and should be free from manipulation. Scherrer (2011) suggested validity is dependent on other statistical and measurement issues, including choice of testing instrument, precision in the estimates, and vertical scaling. Thus, evaluating the validity of the inferences made from VA models is very much tied to resolving other issues considered in this review. Few papers addressed validity in a systematic way, however. Amrein-Beardsley (2008) expressed concern with the lack of evidence of validity of using the estimates from the Education Value-Added Assessment System as indicators of teacher quality.

Hill, Kapitula, and Umland (2011) examined convergent validity of VA estimates by correlating the teacher effect estimates with observation scores. Although there was a correlation between the two sets of scores, some teachers were classified differently using the two approaches, and scores were correlated with student traits. In addition to finding some evidence of convergent validity, the authors also found some evidence of divergent validity. The authors concluded that, although VA scores do seem to contain some information about teacher quality, they probably do not contain enough information to be used on their own. Van de Grift (2009) found that VA models only moderately predicted school equity, and he suggested this result provided evidence against the validity of using the estimates as predictors of school quality.

Strunk, Weinstein, and Makkonen (2014) examined the relationship between VA estimates and observations of teachers, and they found only moderate correlations, suggesting the two approaches may measure similar, but not identical, constructs. The relationship was stronger for teachers in the upper two quintiles of the distribution than for those in the lower two quintiles. The authors found, however, a stronger relationship between 1 year VA estimates and the observations than between 3 year VA estimates and the observations. Thus, 1 year VAMs may be more appropriate for measuring this year’s performance for a teacher. Together, this literature represents a starting point in investigating the validity of making inferences about teacher quality from VA estimates.
Summary

Although several researchers listed or examined other statistical assumptions that may be of concern (e.g., Harris & McCaffrey, 2010; Karl et al., 2013; Reardon & Raudenbush, 2009), the statistical issues presented here represent concerns that have been identified by at least two works in the recent literature. As suggested by Sass et al. (2014), there is plentiful evidence that all statistical assumptions of VA models are violated, and the question becomes one of “to what degree?” As these authors indicate, the absence of random assignment makes the estimation of the extent of actual bias impossible. It appears that the statistical issues that received the most research attention, however—that is, classroom sorting, persistence, and stability—are issues for which creating a framework from which to study them is manageable, even if solving the problem is not. Those issues that received less attention— attribution, school sorting, and validity—are those for which defining the problem is more complex.

Measurement Issues

Although the focus of this review is on VA modeling rather than test properties, many VA researchers have argued that certain measurement problems directly affect the VA estimates or interact with the statistical assumptions required by VA models. As shown in Table 1, theorists frequently had dual concerns with statistical and measurement issues, and the divide between the two has become less stark. Although a few mentioned other measurement issues such as ceiling effects (e.g., Koedel & Betts, 2009), the dominant measurement concerns addressed by VA researchers are discussed below.

Measurement Error

Hanushek and Rivkin (2010) asserted that when VA estimates were adjusted for measurement error, their variance decreased. However, they indicated that measurement error issues were less relevant with multiple years of data or larger sample sizes. Jakubowski (2008) found that adjusting for measurement error altered VA estimates as well as their rankings to a degree that was practically important. Using fixed effects, this problem was especially relevant for schools not near the mean, but using random effects the problem was large for all schools. Papay (2011) explained instability in the rankings of VA estimates across reading tests as largely a result of measurement error and suggested the use of multiple assessments as one way of minimizing this risk. Similarly, Koedel, Leatherman, and Parsons (2012) found that by adjusting for measurement error, VA estimates became more precise to the degree that would be found if teacher class sizes were increased 11% to 17%.

On the other hand, Boyd, Grossman, Lankford, Loeb, and Wyckoff (2008) found differences in the size, but not the rankings, of VA estimates when measurement error was accounted for, and Ferrão and Goldstein (2009) concluded that random school effects were not affected by accounting for measurement error. Ferrão and Goldstein suggested, however, that the issues of measurement error variation both along the score scale and across schools may need further study. Although the literature addressing measurement error was minimal and the conclusions were inconsistent, it suggested that measurement error was not a large
problem within a limited context (ability level, school, test) but may become a larger issue across contexts, suggesting the need for further study.

Equal Interval Scale

An interval, or equal interval, scale is one in which a gain of the same size means the same thing at all points on the scale. This assumption is essential to VA modeling because, if the scale is not equal interval, then it is impossible to make meaningful comparisons of gains across groups. Although some psychometricians have argued that item response theory-based estimates create an equal interval scale, others have contested that claim (Ballou, 2009).

Although many theorists worried about the problems of a nonequal-interval scale (Table 1), only two authors within this literature pool examined the problem. Ballou (2009) concluded that item response theory ability estimates (θs) are probably not equal interval, partially based on data which showed that gains are not the same size from grade to grade. Ballou investigated transformations of the θ scale that would create equal average gains at each grade level and found that these transformations had an enormous impact on VA estimates. Incorrectly assuming that a scale is equal interval may result in inaccurate VA estimates. Ballou attempted using rank-based methodology appropriate for ordinal data, and he obtained effect estimates that did not always categorize teachers the same way as did VA estimates. Although he conceded that, if the θ scale truly was interval, then the ordinal methodology did not fully use all information, he suggested that his ordinal estimates were more defensible. He also admitted that many of the other concerns expressed by researchers about VA modeling may also apply to ordinal methods. Although there is a need for more research, Ballou has presented a unique approach to examining this problem. Briggs and Domingue (2013) reacted to Ballou (2009) by suggesting a simple method for identifying the degree of departure from an interval scale. Using one example, the authors did not find that the departure from an interval scale significantly affected rankings of schools using a VA gain score approach.

Vertical Scaling

Vertical scaling puts scores resulting from nonidentical tests taken across grades onto the same scale so that they are comparable and is necessary when a gain model approach is used. Theorists have expressed concern that the choice of vertical scaling methodology may affect VA estimates (Table 1). However, only two studies within this group of literature addressed the problem. Briggs and Weeks (2009) estimated VA school effects after using eight different combinations of vertical scaling procedures. They found that, although the effects were highly correlated, the precision of the effects depended on the scaling method used. The authors stated that further research needed to address the problem of vertical scaling in the presence of multidimensionality. In 2013, Briggs and Domingue conducted another study of the effects of vertical scaling on VA estimates based on gain scores and found that there was greatest impact when the vertical scaling approach resulted in large changes in variability in student outcomes across grade levels. The authors conclude that vertical scaling is a more important issue when inferences are being made about growth of individual students than when they are made about teachers or schools.
Choice of Testing Instrument

There was also concern among theorists that the choice of testing instrument itself might affect VA estimates (Table 1). Lockwood, McCaffrey, Hamilton, et al. (2007, p. 48) explained the importance of this problem:

If VAM measures are highly sensitive to specific properties of the achievement measures, then educators and policy makers might conclude that VAM measures are too capricious to be used fairly for accountability. On the other hand, if the measures are robust to different measures of the same broad content area, then educators and policymakers might be more confident in their use.

Lockwood et al. used two subtests from the SAT-9 math test for Grades 6 to 8 and calculated VA estimates under a variety of models. The authors discovered that the choice of testing instrument caused greater instability in the VA estimates than did the model choice. The authors also noted that these comparisons were not across publishers and did not take into account fit with the curriculum, implying that the problem may be even larger if tests from multiple publishers were used. McCaffrey et al. (2009), in their article addressing intertemporal stability, found inconsistency in the stability estimates across testing instruments, suggesting a possible interaction between intertemporal and intertest instability that may require further research.

Summary

Although little recent VA research addressed measurement issues, it was a frequent area of concern among theorists (Table 1). It appears, however, that researchers have begun to develop frameworks for analyzing the extent of problems with measurement issues that could be built upon. The more challenging problems seem to be those that have underlying philosophical questions that must be answered. What exactly would equal gains across grades mean? What should be tested? Is there such a thing as comparability across constructs or ability levels?

Conclusion

A review of the most recent VA literature suggests that our knowledge about how well VA models estimate teacher or school contributions to student test scores is still evolving, in spite of at least one author’s assertion that the properties of VA estimates are well known (Corcoran & Goldhaber, 2013). Even when research has offered insight, conclusions need to be validated using different populations and different testing instruments. Across the issues of theorist concern (see Table 1), there were areas of consensus and resolution, but there were also issues that have not been addressed adequately (see Table 3).

Modeling Decisions

A large number of papers in this pool reflected on or directly investigated modeling issues (Table 3), in spite of these issues being somewhat sporadically mentioned as concerns by the theorists (Table 1). Some issues are well understood
statistically, such as the consequences of choosing fixed rather than random
effects. However, how to apply these alternatives within the VA modeling context
is not always clear. Random effects, for example, appear to offer most benefit,
with least added risk of bias, when assignment to teachers or schools is well
understood and can be modeled with covariates (Clarke et al., 2010; Ray et al.,
2009). Understanding this selection mechanism, however, is practically problem-
atic as suggested by the lack of consensus on the addition of covariates to the
model (see Covariate Choice subsection). Similarly, there appears to be consensus
in the literature that school effects cannot be included in models meant to estimate
teacher effects, but a few have developed a potential work-around using school
changers that might be more fully explored (Bacher-Hicks et al., 2014; Chetty
et al., 2014a).

Several other issues related to modeling choice have in common that they
focus on decisions related to model complexity, and there is minimal consensus
regarding how to address these four issues in the recent literature (Table 3). In all
four cases, there is a tension between producing estimates with minimal bias and
creating models that are practically useful. Desired covariates are not always

---

### TABLE 3
*Consensus in the research literature*

| Issue                        | Number of papers | Consensus  |
|------------------------------|------------------|------------|
|                              |                  | Full or near full | Partial | Minimal or none |
| Modeling decisions           |                  |             |          |                |
| Gain vs. covariate adjustment| 7                | X           |          |                |
| Random vs. fixed effects     | 11               | X           |          |                |
| Covariate choice             | 18               | X           |          |                |
| School vs. teacher effects   | 10               | X           |          |                |
| Pretest timing               | 3                | X           |          |                |
| Differential effects         | 7                | X           |          |                |
| Transparency                 | 4                | X           |          |                |
| Statistical assumptions and interpretation |                  |             |          |                |
| Sorting bias                 | 19               | X           |          |                |
| Imprecision and instability  | 19               | X           |          |                |
| Persistence                  | 6                | X           |          |                |
| Attribution                  | 6                | X           |          |                |
| Validity                     | 3                | X           |          |                |
| Measurement                  |                  |             |          |                |
| Equal interval scale         | 2                | X           |          |                |
| Vertical scaling             | 2                | X           |          |                |
| Measurement error            | 6                | X           |          |                |
| Choice of testing instrument | 2                | X           |          |                |

*Note.* See appropriate subsection in review for specific papers’ references.
available (Rothstein, 2009), and even if available, researchers are divided as to whether missing covariates significantly bias estimates (Clarke et al., 2010; Dearden et al., 2011; Jackson, 2012; Johnson et al., 2013; Levine & Painter, 2008; Protik et al., 2013) or have no important impact at all (Harris & McCaffrey, 2010; Jakubowski, 2008; Lenkeit, 2013; Levine & Painter, 2008; Lockwood, McCaffrey, Hamilton, et al., 2007; Schochet & Chiang, 2013). One covariate that all models use, a prior test score, has even been debated in terms of timing, fall of the academic year, or the prior spring. Although evidence suggests accounting for the summer learning loss by including the fall test score affects estimates (Palardy, 2010), adding the fall score increases complexity, places burdens on schools, and creates an opportunity for teachers to game the system (Baker et al., 2010).

Another issue, the choice to model differential teacher effects, is particularly problematic in that it creates complexity in the interpretation of effects, and researchers are divided about whether it makes a difference (Chetty et al., 2014b; Condie et al., 2014; Goldhaber et al., 2013; Reardon & Raudenbush, 2009) or produces unnecessary complications (Jakubowski, 2008; Lockwood & McCaffrey, 2009). In summary, a critical need that has not been adequately resolved by the VA literature is determining what actually belongs in the model.

**Statistical Assumptions and Interpretation**

A large number of papers address each issue in this area, as reflected by Table 3, suggesting a promising focus on the problems that the theorists most frequently identified (Table 1). Of these statistical issues, the potential problems resulting from sorting bias was both the most frequently worried about by theorists, and the most frequently addressed by researchers. At the teacher level, Rothstein’s studies (2007, 2009, 2010, 2015) were the primary source of negative evidence regarding the promise of VA estimates to overcome sorting bias, with most other researchers finding that the estimates were more promising (Altonji & Mansfield, 2014; Bacher-Hicks et al., 2014; Chetty et al., 2014a; Deming, 2014; Guarino et al., 2014; Kane & Staiger, 2008; Koedel & Betts, 2011). At the school level, the evidence was less substantial but more evenly divided. These results suggest the possibility that concerns regarding bias are not always supported, although several suggested that in high-stakes settings, principals might be tempted to sort students to teachers in a way that has a larger effect (Koedel & Betts, 2011; Rothstein, 2009). In addition, there has been little work regarding the interaction of sorting with other issues such as covariate choice or instability.

There appears to be a strong consensus, however, that concerns regarding the potential for lack of persistence or instability of effect estimates are well-founded. There is little agreement, however, as to whether effect instability is due to real differences in teacher quality (over instrument, over time, etc.) or results from bias or imprecision in the estimates themselves. The common solution of averaging estimates (e.g., across years) generally assumes that the underlying trait is stable, but this assumption needs validation.

One challenging issue of interpretation relates to attribution. Although six papers addressed the problem, only half of these were more than conjecture. Finding ways to adequately attribute effect estimates to teachers or school efforts remains one of the most critical issue in VA research, particularly in light of
increasing efforts to use VA estimates for high-stakes purposes, such as teacher evaluation. Validity also remains an underinvestigated issue that is critical if VA estimates are to be used for high-stakes purposes.

**Measurement**

In spite of the frequent mention of measurement issues among theorists (Table 1), measurement was the least-addressed topic in the reviewed research literature (Table 3). Thus, further research is important. Measurement error was the most-researched measurement-related topic, but there was little consensus as to its effect on VA estimates. Half the reviewed papers concluded that measurement error is not big enough to be problematic (Boyd et al., 2008; Ferrão & Goldstein, 2009; Hanushek & Rivkin, 2010), and the other half concluded it does affect the estimates (Jakubowski, 2008; Koedel et al., 2012; Papay, 2011). Furthermore, researchers suggested measurement error affects other issues such as the decision to select random or fixed effects (Jakubowski, 2008), precision (Koedel et al., 2012), and stability across instruments (Papay, 2011). The measurement error problem, if it exists, may make other problems worse, suggesting a real need to come to consensus on this issue. The issues of an equal interval scale and vertical scaling were inadequately addressed with only two papers investigating each topic. Each paper concluded, however, that the scaling issue is potentially relevant to VA estimates, suggesting the critical need for more work in this area.

**Summary**

This literature review suggests we are clearly not at the end of the journey, but we do have a beginning foundation of knowledge. There are many small and large questions about VA that require further research. Three more fundamental concerns about the state of the research arise from this review of the recent literature.

First, the problems that have received the greatest share of the research attention are not always the same as those which theorists suggest are of greatest concern. It appears, for example, that measurement-related issues have not been researched in proportion to concern for them among researchers. In some cases, it appears that the research addresses the questions for which it is most possible to obtain answers, not those for which it is most critical to obtain answers. For example, the ability to attribute a teacher effect to a teacher’s skill, rather than other inputs, is a concern for most theorists (Table 1). No literature in this review, however, suggests a framework or methodology for untangling a teacher’s contribution from confounding influences, such as access to classroom materials, specialist support, or tutoring. Whether this is due to the high costs that might accompany such research or to the complexity of the problem is not clear from this literature review.

Second, most research has focused on individual issues in isolation. The potential for interactions between those issues has barely been acknowledged. One exception was Harris (2009b), who found an interaction between both model form and sorting and the importance of modeling differential effects. He suggested similar interactions may exist elsewhere:
Everson

It is also important to point out that these assumptions may be interrelated so that violating one assumption might compound or offset the impact of violations in other assumptions. Research at present is mainly focused on testing individual assumptions, which is often quite complicated by itself. (Harris, 2009b, p. 331)

An issue that, by itself, is shown to be unimportant may be suddenly critical in the presence of some other VA issue or within a specific context. Measurement issues may be more relevant in the presence of statistical issues and vice versa. Optimal modeling choices, such as decisions about covariate inclusion, may depend on the degree to which statistical issues, such as stability, is relevant to the project.

Third, many of the VA issues we are trying to answer cannot be adequately answered without first stepping back and answering philosophical questions about teacher quality. Primarily, these questions address the definition of a good teacher. Is good teaching a multidimensional construct (Broatch & Lohr, 2012)? Are teacher inputs to nonachievement outcomes just as important (e.g., Ruzek, Domina, Conley, Duncan, & Karabenick, 2015)? Does a good teacher bring all students to a set standard regardless of background or context, or does a good teacher maximum student gains within that background or context? Is a good teacher one who would succeed with any type of student, or one who succeeds with the type of student he or she is actually assigned that year (Condie et al., 2014; Everson et al., 2013; Goldhaber, Cowan, et al., 2013)? Should the school system be rewarded for sorting of students to schools or classrooms in ways that promote good teacher–student matches, or should their VA estimates adjust for this sorting to make teachers comparable? To whom should a teacher be compared? Answers to these questions may color our approach to sorting bias and our decisions about adding previous ability and background variables to VA models.

Another philosophical problem that requires further investigation is related to the comparability of student gains. Is it possible to meaningfully compare gains in Grade 2 with those in Grade 6, even if the same content is assessed? Is it possible to compare gains for students learning English or with learning disabilities with the gains of other students? Is it meaningful to compare gains across testing instruments that do not have identical content coverage? These issues must be answered before measurement issues, such as equal interval scaling, can be approached.

To better understand VA modeling, we must first step back and assess the questions we are asking. Are they the most important ones, or simply the easiest to answer? The list of reasons not to use VA modeling may be long. Holding teachers or schools accountable based on methodologies that do not accurately reflect teacher or school quality could do more harm than good. On the other hand, the list of reasons to use any of the alternatives to VA modeling, if we wish to hold schools or teachers partially accountable for student test scores, could well be longer. This possibility, however, does not excuse us from continuing in the quest to better understand and develop VA modeling methodologies. If we are going to use them, then we must do so well. A first step is to both ask and answer the right questions.
References

Adler, M. (2014). Review of “Measuring the impacts of teachers.” Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/files/terchetty-teachimpacts_0.pdf

Altonji, J. G., & Mansfield, R. K. (2014). Group-average observables as controls for sorting on unobservables when estimating group treatment effects: The case of school and neighborhood effects (Working Paper No. 20781). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w20781

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System. Educational Researcher, 37, 65–75. doi:10.3102/0013189X08316420

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). Validating teacher effect estimates using changes in teacher assignments in Los Angeles (Working Paper No. 20657). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://nber.org/papers/w20657

Baker, E., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., & Linn, R. L. (2010). Problems with the use of student test scores to evaluate teachers (Briefing Paper #278). Washington, DC: Economic Policy Institute. Retrieved from http://epi.3cdn.net/b9667271ee6c154195_t9m6ijj8k.pdf

Ballou, D. (2009). Test scaling and value-added measurement. Education Finance and Policy, 4, 351–383. doi:10.1162/edfp.2009.4.4.351

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. Journal of Educational and Behavioral Statistics, 29, 37–66. doi:10.3102/10769986029001037

Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. Teachers College Record, 116(1), 1–31.

Betebenner, D. W. (2011). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth trajectories/projections. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from http://www.nj.gov/education/njsmart/performance/SGP_Technical_Overview.pdf

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). Overview of measuring effect sizes: The effect of measurement error (CALDER Working Paper No. 19). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.caldercenter.org/sites/default/files/1001257_measuring_effect_sizes.pdf

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. Journal of Educational and Behavioral Statistics, 38, 551–576. doi:10.3102/1076998613508317

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. Education Finance and Policy, 4, 384–414. doi:10.1162/edfp.2009.4.4.384

Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level value-added. Journal of Educational and Behavioral Statistics, 36, 616–637. doi:10.3102/1076998610396887

Broatch, J., & Lohr, S. (2012). Multidimensional assessment of value added by teachers to real-world outcomes. Journal of Educational and Behavioral Statistics, 37, 256–277. doi:10.3102/1076998610396900

63
Everson

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. Journal of Educational and Behavioral Statistics, 39, 333–367. doi:10.3102/1076998614547576

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. American Economic Review, 104, 2593–2632. doi:10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. American Economic Review, 104, 2633–2679. doi:10.1257/aer.104.9.2633

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014c). Response to Adler (2014) review of “Measuring the impacts of teachers.” Retrieved from http://www.rajchetty.com/chettyfiles/Adler_response.pdf

Clarke, P., Crawford, C., Steele, F., & Vignoles, A. (2010). The choice between fixed and random effects models: Some considerations for educational research (IZA DP No. 5287). Bonn, Germany: Institute for the Study of Labor. Retrieved from http://ftp.iza.org/dp5287.pdf

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. Economics of Education Review, 40, 76–92. doi:10.1016/j.econedurev.2013.11.009

Corcoran, S., & Goldhaber, D. (2013). Value added and its uses: Where you stand depends on where you sit. Education Finance and Policy, 8, 418–434. doi:10.1162/EDFP_a_00104

Dearden, L., Miranda, A., & Rabe-Hesketh, S. (2011). Measuring school value added with administrative data: The problem of missing variables. Fiscal Studies, 32, 263–278. doi:10.1111/j.1475-5890.2011.00136.x

Deming, D. J. (2014). Using school choice lotteries to test measures of school effectiveness. American Economic Review, 104, 406–411. doi:10.1257/aer.104.5.406

Dieterle, S., Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). How do principals assign students to teachers? Finding evidence in administrative data and the implications for value added. Journal of Policy Analysis and Management, 34, 32–58. doi:10.1002/pam.21781

Eckert, J. M., & Dabrowski, J. (2010). Should value-added measures be used for performance pay? Phi Delta Kappan, 91, 88–92. doi:10.1177/003172171009100821

Everson, K. C., Feinauer, D., & Sudweeks, R. (2013). Rethinking teacher evaluation: A conversation about statistical inferences and value-added models. Harvard Educational Review, 83, 349–370. doi:10.17763/haer.83.2.m32hk8q851u752h0

Ferrão, M. E. (2012). On the stability of value added indicators. Quality & Quantity, 46, 627–637. doi:10.1007/s11135-010-9417-6

Ferrão, M. E., & Goldstein, H. (2009). Adjusting for measurement error in the value added model: Evidence from Portugal. Quality & Quantity, 43, 951–963. doi:10.1007/s11135-008-9171-1

Goldhaber, D., & Chaplin, D. (2012). Assessing the Rothstein test: Does it really show teacher value-added models are biased? Washington, DC: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/~/media/publications/pdfs/education/rothstein_wp.pdf

Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance across subjects. Economics of Education Review, 36, 216–228. doi:10.1016/j.econedurev.2013.06.010
Goldhaber, D., Goldschmidt, P., & Tseng, F. (2013). Teacher value-added at the high school level: Different models, different answers? *Educational Evaluation and Policy Analysis, 35*, 220–236. doi:10.3102/0162373712466938

Goldhaber, D., & Hansen, M. (2008). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.urban.org/sites/default/files/alfresco/publication-pdfs/1001265-Assessing-the-Potential-of-Using-Value-Added-Estimates-of-Teacher-Job-Performance-for-Making-Tenure-Decisions.PDF

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica, 80*, 589–612. doi:10.1111/ecca.12002

Gorard, S. (2008). The value-added of primary schools: What is it really measuring? *Educational Review, 60*, 179–185. doi:10.1080/00131910801934185

Gorard, S. (2011). Now you see it, now you don’t: School effectiveness as conjuring? *Research in Education, 86*, 39–45. doi:10.7227/RIE.86.4

Gruino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An evaluation of empirical Baye’s estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics, 40*, 190–222. doi:10.3102/1076998615574771

Gruino, C. M., Reckase, M., Stacy, B., & Wooldridge, J. (2014). Evaluating specification tests in the context of value-added estimation (Working Paper No. 38). East Lansing: Michigan State University. Retrieved from http://education.msu.edu/epc/publications/documents/WP38EvaluatingSpecificationTests.pdf

Gruino, C. M., Reckase, M., & Wooldridge, J. (2012). *Can value-added measures of teacher performance be trusted* (Discussion Paper No. 6602)? Bonn, Germany: Institute for the Study of Labor. Retrieved from http://ftp.iza.org/dp6602.pdf

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*, 267–271. doi:10.1257/aer.100.2.267

Harris, D. N. (2009a). Teacher value-added: Don’t end the search before it starts. *Journal of Policy Analysis and Management, 28*, 693–699. doi:10.1002/pam.20464

Harris, D. N. (2009b). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy, 4*, 319–350. doi:10.1162/edfp.2009.4.4.319

Harris, D. N. (2010). Clear away the smoke and mirrors of value-added. *Phi Delta Kappan, 91*, 66–69. doi:10.1177/003172171009100813

Harris, D. N., & McCaffrey, D. F. (2010). Value-added: Assessing teachers’ contributions to student achievement. In M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality* (pp. 251–282). San Francisco, CA: Jossey-Bass.

He, Q., & Tymms, P. (2014). The principal axis approach to value-added calculation. *Educational Research and Evaluation, 20*, 25–43. doi:10.1080/13803611.2013.860036

Henry, G. T., & Rose, R. A. (2015). Are value-added models good enough for teacher evaluations? Assessing commonly used models with simulated and actual data. *Investigaciones de Economía de La Educación, 9*, 383–405. Retrieved from http://repec.economicsofeducation.com/2014valencia/09-20.pdf

Herrmann, M. A., Walsh, E., Isenberg, E., & Resch, A. (2013). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels (Working Paper No. 17). Washington, DC: Mathematica Policy Research.
Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28*, 700–709. doi:10.1002/pam.20463

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*, 794–831. doi:10.3102/0002831210387916

Hock, H., & Isenberg, E. (2012). *Methods for accounting for co-teaching in value-added models* (Working Paper No. 6). Washington, DC: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/~/media/publications/PDFS/education/acctco-teaching_wp.pdf

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*, 794–831. doi:10.3102/0002831210387916

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy, 4*, 520–536. doi:10.1162/efdp.2009.4.4.520

Jackson, C. K. (2009). Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics, 27*, 213–256. doi:10.1086/599334

Jackson, C. K. (2012). *Teacher quality at the high-school level: The importance of accounting for tracks* (Working Paper No. 17722). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w17722.pdf

Jacob, B. A., Lefgren, L., & Sims, D. (2008). *The persistence of teacher-induced learning gains* (NBER Working Paper No. 14065). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://econpapers.repec.org/RePEc:nbr:nberwo:14065

Jakubowski, M. (2008). *Implementing value-added models of school assessment* (EUI RSCAS Working Paper No. 2008/06). Florence, Italy: European University Institute. Retrieved from http://ideas.repec.org/p/rscrsceu/2008-06.html

Jennings, J. L., & Corcoran, S. P. (2009). Beware of geeks bearing formulas. *Phi Delta Kappan, 90*, 635–639. doi:10.1177/003172170909009096

Johnson, M., Lipscomb, S., & Gill, B. (2013). Sensitivity of teacher value-added estimates to student and peer control variables (Working Paper No. 25). Washington, DC: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/~/media/publications/PDFS/education/value-added_estimates_wp.pdf

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14607

Karl, A. T., Yang, Y., & Lohr, S. L. (2013). A correlated random effects model for non-ignorable missing data in value-added assessment of teacher effects. *Journal of Educational and Behavioral Statistics, 38*, 577–603. doi:10.3102/1076998613494819

Kelly, A., & Downey, C. (2010). Value-added measures for schools in England: Looking inside the “black box” of complex metrics. *Educational Assessment, Evaluation and Accountability, 22*, 181–198. doi:10.1007/s11092-010-9100-4

Kersting, N. B., Chen, M.-K., & Stigler, J. W. (2013). Value-added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications
on the stability of teacher value-added scores. *Education Policy Analysis Archives*, 21(7), 1–39. doi:10.14507/epaa.v21n7.2013

Kinsler, J. (2012a). Assessing Rothstein’s critique of teacher value-added models. *Quantitative Economics*, 3, 333–362. doi:10.3982/QE132

Kinsler, J. (2012b). Beyond levels and growth: Estimating teacher value-added and its persistence. *Journal of Human Resources*, 47, 722–753. doi:10.1353/jhr.2012.0023

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review*, 28, 682–692. doi:10.1016/j.econedurev.2009.02.003

Koedel, C., & Betts, J. R. (2009). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy*, 5, 54–81. doi:10.1162/edfp.2009.5.1.5104

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, 18–42. doi:10.1162/EDFP_a_00027

Koedel, C., Leatherman, R., & Parsons, E. (2012). Test measurement error and inference from value-added models (Working Paper). Columbia: University of Missouri. Retrieved from https://economics.missouri.edu/working-papers/2012/wp1201_koedel.pdf

Kyriakides, L., & Creemers, B. P. M. (2008). A longitudinal study on the stability over time of school and teacher effects on student outcomes. *Oxford Review of Education*, 34, 521–545. doi:10.1080/03054980701782064

Leckie, G. (2009). The complexity of school and neighborhood effects and movements of pupils on school differences in models of educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 537–554. doi:10.1111/j.1467-985X.2008.00577.x

Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis*, 34, 109–121. doi:10.3102/0162373711422377

Lenkeit, J. (2013). Effectiveness measures for cross-sectional studies: A comparison of value-added models and contextualised attainment models. *School Effectiveness and School Improvement*, 24, 39–63. doi:10.1080/09243453.2012.680892

Levine, D. I., & Painter, G. (2008). Are measured school effects just sorting? *Economics of Education Review*, 27, 460–470. doi:10.1016/j.econedurev.2007.01.004

Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies*, 40, 699–711. doi:10.1080/00220270802105729

Lockwood, J. R., & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, 4, 439–467. doi:10.1162/edfp.2009.4.4.439

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44, 47–67. doi:10.1111/j.1745-3984.2007.00026.x

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, 32, 125–150. doi:10.3102/1076998606298039

Loeb, S., Soland, J., & Fox, L. (2014). Is a good teacher a good teacher for all? Comparing value-added of teachers with their English learners and non-English learners. *Educational Evaluation and Policy Analysis*, 36, 457–475. doi:10.3102/0162373714527788
Lopez-Martin, E., Kuosmanen, T., & Gaviria, J. L. (2014). Linear and nonlinear growth models for value-added assessment: An application to Spanish primary and secondary schools’ progress in reading comprehension. *Educational Assessment, Evaluation and Accountability, 26*, 361–391. doi:10.1007/s11092-014-9194-1

Manzi, J., Martin, E. S., & Van Bellegem, S. (2010). *School system evaluation by value-added analysis under endogeneity* (CORE Discussion Paper 2010/46). Louvain-la-Neuve, Belgium: Center for Operations Research and Econometrics. Retrieved from https://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2010_46web.pdf

Martineau, J. A. (2010). The validity of value-added models: An allegory. *Phi Delta Kappan, 91*, 64–67. doi:10.1177/003172171009100713

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*, 572–606. doi:10.1162/edfp.2009.4.4.572

Misco, T. (2008). Was that a result of my teaching? A brief exploration of value-added assessment. *Clearing House: A Journal of Educational Strategies, Issues and Ideas, 82*, 11–14. doi:10.3200/TCHS.82.1.11-14

National Conference of State Legislatures. (2010). *Education bill tracking database*. Washington, DC: Author. Retrieved from http://www.ncsl.org/IssuesResearch/Education/EducationBillTrackingDatabase/tabid/12913/Default.aspx

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*, 1–24.

Organization for Economic Cooperation and Development. (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. Paris, France: Author. Retrieved from http://www.oecd.org/edu/school/measuringimprovementsinlearningoutcomesbestpracticestoassessthevalue-addedofschools.htm

Palardy, G. J. (2010). The multilevel crossed random effects growth model for estimating teacher and school effects: Issues and extensions. *Educational and Psychological Measurement, 70*, 401–419. doi:10.1177/0013164409355693

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*, 163–193. doi:10.3102/0002831210362589

Protik, A., Walsh, E., Resch, A., Isenberg, E., & Kopa, E. (2013). *Does tracking of students bias value-added estimates for teachers?* (Working Paper No. 15). Oakland, CA: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/~media/publications/PDFs/education/tracking_students_value_added_wp.pdf

Ray, A., McCormack, T., & Evans, H. (2009). Value added in English schools. *Education Finance and Policy, 4*, 415–438. doi:10.1162/edfp.2009.4.4.415

Ready, R. (2013). Associations between student achievement and student learning: Implications for value-added school accountability models. *Educational Policy, 27*, 92–120. doi:10.1177/0895904814429289

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*, 492–519. doi:10.1162/edfp.2009.4.4.492

Rothstein, J. (2007). *Do value-added models add value? Tracking, fixed effects, and causal inference* (CEPS Working Paper No. 159). Princeton, NJ: Center for Economic Policy Studies. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.188.8444
Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*, 537–571. doi:10.1162/edfp.2009.4.4.537

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*, 175–214. doi:10.1162/qjec.2010.125.1.175

Rothstein, J. (2015, October). *Revisiting the impacts of teachers* (UC-Berkeley Working Paper). Retrieved from http://econ.berkeley.edu/~jrothst/workingpapers/rothstein_cfr.pdf

Ruzek, E. A., Domina, T., Conley, A. M., Duncan, G. J., & Karabenick, S. A. (2015). Using value-added models to measure teacher effects on students’ motivation and achievement. *Journal of Early Adolescence, 35*, 852–882. doi:10.1177/0272431614525260

Sanders, W. L. (1998). Value-added assessment. *School Administrator, 55*, 24–77.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299–311. doi:10.1007/BF00973726

Sanders, W. L., Wright, P., & Langevin, W. E. (2008). Do teacher effect estimates persist when teachers move to schools with different socioeconomic environment (Working Paper 2008-20)? Nashville, TN: The National Center on Performance Initiatives, Vanderbilt University. Retrieved from http://education.ohio.gov/getattachment/Topics/Data/Accountability-Resources/Value-Added-Resources/Do-Teacher-Effect-Estimates-Persist-When-Teachers-Move-to-Schools-with-Different-Socioeconomic-Environments.pdf.aspx

Sass, T. R. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy* (Issue Brief No. 4). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://www.urban.org/publications/1001266.html

Sass, T. R., Semykina, A., & Harris, D. N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review, 38*, 9–23. doi:10.1016/j.econedurev.2013.10.003

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin, 95*, 122–140. doi:10.1177/0192636511410052

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*, 142–171. doi:10.3102/1076998611432174

Snijders, T. A. B. (2005). Fixed and random effects. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 664–665). Chichester, UK: Wiley. doi:10.1002/0470013192.bsa234

Stacy, B. (2014). *Ranking teachers when teacher value-added is heterogeneous across students* (ECONSTOR Preprints). Kiel, Germany: German National Library of Economics. Retrieved from http://econstor.eu/bitstream/10419/104743/1/wp_teacher_mean_var.pdf

Stacy, B., Guarino, C., Reckase, M., & Wooldridge, J. (2013). *Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve* (Discussion Paper No. 7676)? Bonn, Germany: Institute for the Study of Labor. Retrieved from http://ftp.iza.org/dp7676.pdf

Strunk, K., Weinstein, T., & Makkonen, R. (2014). Sorting out the signal: Do multiple measures of teachers’ effectiveness provide consistent information to teachers and
Everson

principals? Education Policy Analysis Archives, 22(100), 1–38. doi:10.14507/epaa.v22.1590

Timmermans, A. C., Doolaard, S., & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. School Effectiveness and School Improvement, 22, 393–413. doi:10.09243453.2011.590704

Timmermans, A. C., Snijders, T. A. B., & Bosker, R. J. (2012). In search of value added in the case of complex school effects. Educational and Psychological Measurement, 73, 210–228. doi:10.1177/0013164412460392

Van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. School Effectiveness and School Improvement, 20, 269–285. doi:10.1080/09243450902883946

Zamarro, G., Engberg, J., Saavedra, J. E., & Steele, J. (2014). Disentangling disadvantage: Can we distinguish good teaching from classroom composition (Paper No. 2014-001)? Retrieved from http://www.researchgate.net/publication/259674242_Disentangling_Disadvantage_Can_We_Distinguish_Good_Teaching_from_Classroom_Composition

Author

KIMBERLEE C. EVERSON is an assistant professor of quantitative methods in the Department of Educational Administration, Leadership, and Research, Western Kentucky University, 1906 College Heights Blvd, Bowling Green, KY 42101, USA; email: kimberlee.everson@wku.edu. She is a graduate of Brigham Young University’s doctoral program in Educational Inquiry, Measurement, and Evaluation. Her research focuses on statistical and psychometric issues related to student growth and teacher evaluation. In particular, she studies quasi-experimental methods, value-added modeling, and teacher beliefs related to student outcomes-based accountability. She also has experience as a mathematics educator, focusing on statistics education and adult developmental mathematics.