Value-Added Modeling: A Review

Cory Koedel[a]
Kata Mihaly[b]
Jonah E. Rockoff[c]

January 2015

This article reviews the literature on teacher value-added. Although value-added models have been used to measure the contributions of numerous inputs to educational production, their application toward identifying the contributions of individual teachers has been particularly contentious. Our review covers articles on topics ranging from technical aspects of model design to the role that value-added can play in informing teacher evaluations in practice, highlighting areas of consensus and disagreement in the literature. Although a broad spectrum of views is reflected in available research, along a number of important dimensions the literature is converging on a widely-accepted set of facts.

JEL Codes: I20, J45, M52
Keywords: Value-added models, VAM, teacher evaluation, education production

Acknowledgement
The authors gratefully acknowledge useful comments from Dan McCaffrey. The usual disclaimers apply.

[a] Department of Economics and Truman School of Public Affairs at the University of Missouri-Columbia
[b] RAND Corporation
[c] Graduate School of Business at Columbia University and National Bureau of Economic Research

This is a preprint version of the article. The final version is forthcoming at *Economics of Education Review*.

1. Introduction

Value-added modeling has become a key tool for applied researchers interested in understanding educational production. The "value-added" terminology is borrowed from the long-standing production literature in economics – in that literature, it refers to the amount by which the value of an article is increased at each stage of the production process. In education-based applications, the idea is that we can identify each student's human capital accumulation up to some point, say by the conclusion of period *t-1*, and then estimate the value-added to human capital of inputs applied during period *t*.

Value-added models (VAMs) have been used to estimate value-added to student achievement for a variety of educational inputs. The most controversial application of VAMs has been to estimate the effects of individual teachers. Accordingly, this review focuses on the literature surrounding teacher-level VAMs.[1] The attention on teachers is motivated by the consistent finding in research that teachers vary dramatically in their effectiveness as measured by value-added (Hanushek and Rivkin, 2010). In addition to influencing students' short-term academic success, access to high-value-added teachers has also been shown to positively affect later-life outcomes for students including wages, college attendance, and teenage childbearing (Chetty, Friedman and Rockoff, 2014b). The importance of access to effective teaching for students in K-12 schools implies high stakes for personnel policies in public education. Chetty, Friedman and Rockoff (2014b) and Hanushek (2011) monetize the gains that would come from improving the quality of the teaching workforce – using value-added-based evidence – and conclude that the gains would be substantial.

[1] Other applications of value-added include evaluations of teacher professional development and coaching programs (Biancarosa, Byrk and Dexter, 2010; Harris and Sass, 2011), teacher training programs (Goldhaber, Liddle and Theobald, 2013; Koedel et al., forthcoming; Mihaly et al., 2013b), reading reform programs (Betts, Zau and King, 2005) and school choice (Betts and Tang, 2008), among others.

The controversy surrounding teacher value-added stems largely from its application in public policy, and in particular the use of value-added to help inform teacher evaluations. Critics of using value-added in this capacity raise a number of concerns, of which the most prominent are (1) value-added estimates may be biased (Baker et al., 2010, Paufler and Amrein-Beardsley, 2014; Rothstein, 2009, 2010), and (2) value-added estimates seem too unstable to be used for high-stakes personnel decisions (Baker et al., 2010; Newton et al., 2010). Rothstein (2015) also raises the general point that the labor-supply response to more rigorous teacher evaluations merits careful attention in the design of evaluation policies. We discuss these and other issues over the course of the review.

The remainder of the paper is organized as follows. Section 2 provides background information on value-added and covers the literature on model-specification issues. Section 3 reviews research on the central questions of bias and stability in estimates of teacher value-added. Section 4 combines the information from Sections 2 and 3 in order to highlight areas of emerging consensus with regard to model design. Section 5 documents key empirical facts about value-added that have been established by the literature. Section 6 discusses research on the uses of teacher value-added in education policy. Section 7 concludes.

2. Model Background and Specification

2.1 Background

Student achievement depends on input from teachers and other factors. Value-added modeling is a tool that researchers have used in their efforts to separate out teachers' individual contributions. In practice, most studies specify linear value-added models in an *ad hoc* fashion, but under some conditions these models can be formally derived from the following cumulative achievement function, taken from Todd and Wolpin (2003) and rooted in the larger education production literature (Ben-Porath, 1967; Hanushek, 1979):

$$A_{it} = A_t[X_i(t), F_i(t), S_i(t), \alpha_{i0}, \varepsilon_{it}] \quad (1)$$

Equation (1) describes the achievement level for student *i* at time *t* (A_{it}) as the end product of a cumulative set of inputs, where $X_i(t)$, $F_i(t)$ and $S_i(t)$ represent the history of individual, family and school inputs for student *i* through year *t*, α_{i0} represents student *i*'s initial ability endowment and ε_{it} is an idiosyncratic error. The intuitive idea behind the value-added approach is that to a rough approximation, prior achievement can be used as a sufficient statistic for the history of prior inputs and, in some models, the ability endowment. This facilitates estimation of the marginal contribution of contemporaneous inputs, including teachers, using prior achievement as a key conditioning variable.

In deriving the conditions that formally link typically-estimated VAMs to the cumulative achievement function, Todd and Wolpin (2003) express skepticism that they will be met. Their skepticism is warranted for a number of reasons. As one example, in the structural model parental inputs can respond to teacher assignments, allowing for increased (decreased) parental inputs that are complements (substitutes) for higher teacher quality. VAM researchers cannot measure and thus cannot control for parental inputs, which means that unlike in the structural model, value-added estimates of teacher quality are inclusive of any parental-input adjustments. More generally, the model shown in equation (1) is flexible along a number of dimensions in ways that are difficult to emulate in practical modeling applications (for further discussion see Sass, Semykina and Harris, 2014).

Sass, Semykina and Harris (2014) directly test the conditions linking VAMs to the cumulative achievement function and confirm the skepticism of Todd and Wolpin (2003), showing that they are not met for a number of common VAM specifications. The tests performed by Sass, Semykina and Harris (2014) give us some indication of what value-added *is not*. In particular, they show that the

parameters estimated from a range of commonly estimated value-added models do not have a structural interpretation. But this says little about the informational value contained by value-added measures. Indeed, Sass, Semykina and Harris (2014) note that "failure of the underlying [structural] assumptions does not necessarily mean that value-added models fail to accurately classify teacher performance" (p. 10).[2] The extent to which measures of teacher value-added provide useful information about teacher performance is ultimately an empirical question, and it is this question that is at the heart of value-added research literature.

2.2 Specification and Estimation Issues

A wide variety of value-added models have been estimated in the literature to date. In this section we discuss key specification and estimation issues. To lay the groundwork for our discussion consider the following linear VAM:

$$Y_{isjt} = \beta_0 + Y_{isjt-1}\beta_1 + X_{isjt}\beta_2 + S_{isjt}\beta_3 + T_{isjt}\theta + \varepsilon_{isjt} \quad (2)$$

In equation (2), Y_{isjt} is a test score for student *i* at school *s* with teacher *j* in year *t*, X_{isjt} is a vector of student characteristics, S_{isjt} is a vector of school and/or classroom characteristics, T_{isjt} is a vector of teacher indicator variables and ε_{isjt} is the idiosyncratic error term. The precise set of conditioning variables in the *X*-vector varies across studies. The controls that are typically available in district and state administrative datasets include student race, gender, free/reduced-price lunch status, language status, special-education status, mobility status (e.g., school changer), and parental education, or some subset therein (examples of studies from different locales that use control variables from this list include Aaronson, Barrow and Sander, 2007; Chetty, Friedman and Rockoff, 2014a; Goldhaber and Hansen, 2013; Kane et al., 2013; Koedel and Betts, 2011; Sass et al., 2012). School and

[2] Guarino, Reckase and Wooldridge (2015) perform simulations that support this point. Their findings indicate that VAM estimators tailored toward structural modeling considerations can perform poorly because they focus attention away from more important issues.

classroom characteristics in the *S*-vector are often constructed as aggregates of the student-level variables (including prior achievement). The parameters that are meant to capture teacher value added are contained in the vector θ. In most studies, teacher effects are specified as fixed rather than random effects.[3]

Equation (2) is written as a "lagged-score" VAM. An alternative, restricted version of the model where the coefficient on the lagged test score is set to unity ($\beta_1 = 1$) is referred to as a "gain-score" VAM. The "gain-score" terminology comes from the fact that the lagged-score term with the restricted coefficient can be moved to the left-hand side of the equation and the model becomes a model of test score gains.[4]

Following Aaronson, Barrow and Sander (2007), the error term in equation (2) can be expanded as $\varepsilon_{isjt} = \lambda_i + \pi_s + e_{isjt}$. This formulation allows for roles of fixed individual ability (λ_i) and school quality (π_s) in determining student achievement growth.[5] Based on this formulation, and assuming for simplicity that all students are assigned to a single teacher at a single school, error in the estimated effect for teacher *j* at school *s* with class size N_j in the absence of direct controls for school and student fixed effects can be written as:

$$E_{sj} = \pi_s + \frac{1}{N_j}\sum_{i=1}^{N_j}\lambda_i + \frac{1}{N_j}\sum_{i=1}^{N_j}e_{ist} \qquad (3)$$

The first two terms in equation (3) represent the otherwise unaccounted for roles of fixed school and student attributes. The third term, although zero in expectation, can be problematic in cases

[3] Examples of studies that specify teacher effects as random include Corcoran, Jennings and Beveridge (2011), Konstantopoulos and Chung (2011), Nye, Konstantopoulos and Hedges (2004), and Papay (2011). Because standard software estimates teacher fixed effects relative to an arbitrary holdout teacher, and the standard errors for the other teachers will be sensitive to which holdout teacher is selected, some studies have estimated VAMs using a sum-to-zero constraint on the teacher effects (Jacob and Lefgren, 2008; Mihaly et al., 2010; Goldhaber, Cowan and Walch, 2013).

[4] In lagged-score VAMs the coefficient on the lagged-score coefficient, unadjusted for measurement error, is typically in the range of 0.6 to 0.8 (e.g., see Andrabi et al., 2011; Rothstein, 2009).

[5] Also see Ishii and Rivkin (2009). The depiction of the error term could be further expanded to allow for separate classroom effects, perhaps driven by unobserved peer dynamics, which would be identifiable in teacher-level VAMs if teachers are observed in more than one classroom.

where N_j is small, which is common for teacher-level analyses (Aaronson, Barrow and Sander, 2007; also see Kane and Staiger, 2002). A number of value-added studies have incorporated student and/or school fixed effects into variants of equation (2) due to concerns about bias from non-random student-teacher sorting (e.g., see Aaronson, Barrow and Sander, 2007; Koedel and Betts, 2011; McCaffrey et al., 2004, 2009; Rothstein, 2010; Sass et al., 2012). These concerns are based in part on empirical evidence showing that students are clearly not randomly assigned to teachers, even within schools (Aaronson, Barrow and Sander, 2007; Clotfelter, Ladd and Vigdor, 2006; Jackson, 2014; Koedel and Betts, 2011; Paufler and Amrein-Beardsley, 2014; Rothstein, 2009, 2010).

An alternative structure to the standard, one-step VAM shown in equation (2) is the two-step VAM, or "average residuals" VAM. The two-step VAM uses the same information as equation (2) but performs the estimation in two steps:

$$Y_{isjt} = \alpha_0 + Y_{isjt-1}\alpha_1 + X_{isjt}\alpha_2 + S_{isjt}\alpha_3 + \eta_{isjt} \tag{4}$$

$$\eta_{isjt} = T_{isjt}\gamma + e_{isjt} \tag{5}$$

The vector γ in equation (5) contains the value-added estimates for teachers. Although most VAMs that have been estimated in the research literature to date use the one-step modeling structure, several recent studies use the two-step approach (Ehlert et al., 2013; Kane, Rockoff and Staiger, 2008; Kane and Staiger, 2008; Kane et al., 2013).

One practical benefit to the two-step approach is that it is less computationally demanding. Putting computational issues aside, Ehlert et al. (forthcoming, 2014) discuss several other factors that influence the choice of modeling structure. One set of factors is related to identification, and in particular identification of the coefficients associated with the student and school/classroom characteristics. The identification tradeoff between the one-step and two-step VAMs can be summarized as follows: the one-step VAM has the potential to "under-correct" for context because

the parameters associated with the control variables may be attenuated (β_1, β_2 and β_3 in equation 2), while the two-step VAM has the potential to "over-correct" for context by attributing correlations between teacher quality and the control variables to the control-variable coefficients (α_1, α_2 and α_3 in equation 4). A second set of factors is related to policy objectives. Ehlert et al. (forthcoming, 2014) argue that the two-step model is better suited for achieving key policy objectives in teacher evaluation systems, including the establishment of an incentive structure that maximizes teacher effort.

It is common in research and policy applications to shrink estimates of teacher value-added toward a common Bayesian prior (Herrmann et al., 2013). In practice, the prior is specified as average teacher value-added. The weight applied to the prior for an individual teacher is an increasing function of the imprecision with which that teacher's value-added is estimated. The following formula describes empirical shrinkage:[6]

$$\hat{\theta}_j^{EB} = a_j * \hat{\theta}_j + (1 - a_j) * \bar{\theta} \tag{6}$$

In (6), $\hat{\theta}_j^{EB}$ is the shrunken value-added estimate for teacher *j*, $\hat{\theta}_j$ is the unshrunken estimate, $\bar{\theta}$ is average teacher value-added, and $a_j = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\lambda}_j}$, where $\hat{\sigma}^2$ is the estimated variance of teacher value-added (after correcting for estimation error) and $\hat{\lambda}_j$ is the estimated error variance of $\hat{\theta}_j$ (e.g., the squared standard error).[7]

The benefit of the shrinkage procedure is that it produces estimates of teacher value-added for which the estimation-error variance is reduced through the dependence on the stable prior. This benefit is particularly valuable in applications where researchers use teacher value-added as an

[6] A detailed discussion of shrinkage estimators can be found in Morris (1983).

[7] There are several ways to estimate σ^2 – e.g., see Aaronson, Barrow and Sander (2007), Chetty, Friedman and Rockoff (2014a), Herrmann et al. (2013), Kane, Rockoff and Staiger (2008), and Koedel (2009).

explanatory variable (e.g., Chetty, Friedman and Rockoff, 2014a/b; Harris and Sass, 2014; Jacob and Lefgren, 2008; Kane and Staiger, 2008; Kane et al., 2013), in which case it reduces attenuation bias. The cost of shrinkage is that the weight on the prior introduces a bias in estimates of teacher value-added. The trend in the literature is clearly toward the increased use of shrunken value-added estimates.[8]

Additional issues related to model specification that have been covered in the literature include the selection of an appropriate set of covariates and accounting for test measurement error. Areas of consensus in the literature regarding the specification issues we have discussed thus far, and these additional issues, have been driven largely by what we know about bias and stability of the estimates that come from different models. It is to these issues that we now turn. In Section 4 we circle back to the model-specification question.

3. Bias and Stability of Estimated Teacher Value-Added

3.1 Bias

A persistent area of inquiry among value-added researchers has been the extent to which estimates from standard models are biased. One consequence of biased value-added estimates is that they would likely lead to an overstatement of the importance of variation in teacher quality in determining student outcomes.[9] Bias would also lead to errors in supplementary analyses that aim to identify the factors that contribute to and align with effective teaching as measured by value-added. In policy applications, a key concern is that if value-added estimates are biased, individual teachers would be held accountable in their evaluations for factors that are outside of their control.

Teacher value-added is estimated using observational data. Thus, causal inference hinges on the assumption that student assignments to teachers (treatments) are based on observables. As with

[8] That said, Guarino et al. (2014) report that shrinkage procedures do not substantially boost the accuracy of estimated teacher value-added. Similarly, the results from Herrmann (2013) imply that the benefits from shrinkage are limited.
[9] The effect of the bias would depend on its direction. The text here presumes a scenario with positive selection bias – i.e., where more effective teachers are assigned to students with higher expected growth.

any selection-on-observables model, the extent to which estimates of teacher value-added are biased depends fundamentally on the degree to which students are sorted to teachers along dimensions that are not observed by the econometrician. Bias will be reduced and/or mitigated when (a) non-random student-teacher sorting is limited and/or (b) sufficiently rich information is available such that the model can condition on the factors that drive the non-random sorting (or factors that are highly correlated with the factors that drive the non-random sorting).

The evidence on bias presented in this section is based on examinations of math and reading teachers in grades 4-8. This focus is driven in large part by data availability – the structure of most standardized testing regimes is such that annual testing occurs in grades 3-8 in math and reading. One direction in which extending the lessons from the current evidence base may be challenging is into high school. High school students are likely to be more-strictly sorted across classrooms, which will make it more difficult to construct VAMs, although not impossible.[10]

Paufler and Amrein-Beardsley (2014) raise concerns about the potential for non-random student-teacher assignments to bias estimates of teacher value-added. They survey principals in Arizona elementary schools and report that a variety of factors influence students' classroom assignments. Many of the factors that they identify are not accounted for, at least directly, in typically-estimated VAMs (e.g., students' interactions with teachers and other students, parental preferences, etc.). The survey results are consistent with previous empirical evidence showing that student-teacher sorting is not random (Aaronson, Barrow and Sander, 2007; Clotfelter, Ladd and Vigdor, 2006; Jackson, 2014; Koedel and Betts, 2011; Rothstein, 2009, 2010), although the survey results are more illuminating given the authors' focus on mechanisms. Paufler and Amrein-Beardsley (2014) interpret their findings to indicate that "the purposeful (nonrandom) assignment of students into classrooms biases value-added estimates" (p. 356). However, no direct tests for bias were

[10] Studies that estimate value added for high school teachers include Aaronson, Barrow and Sander (2007), Jackson (2014), Koedel (2009) and Mansfield (forthcoming).

performed in their study, nor were any value-added models actually estimated. Although Paufler and Amrein-Beardsley (2014) carefully document the dimensions of student sorting in elementary schools, what is not clear from their study is how well commonly-available control variables in VAMs provide sufficient proxy information.

Rothstein (2010) offers what has perhaps been the highest-profile criticism of the value-added approach (also see Rothstein, 2009). He examines the extent to which future teacher assignments predict students' previous test-score growth conditional on standard controls using several commonly estimated value-added models (a gain-score model, a lagged-score model, and a gain-score model with student fixed effects). Rothstein (2010) finds that future teachers appear to have large "effects" on previous achievement growth. Because future teachers cannot possibly cause student achievement in prior years, he interprets his results as evidence that the identifying assumptions in the VAMs he considers are violated. He writes that his finding of non-zero effects for future teachers implies that "estimates of teachers' effects based on these models [VAMs] cannot be interpreted as causal" (p. 210). However, several subsequent studies raise concerns about the construction of Rothstein's tests and how he interprets his results. Perhaps most notably, Goldhaber and Chaplin (2015) show that his tests reject VAMs even when there is no bias in estimated teacher effects (also see Guarino et al., 2015). In addition, Kinsler (2012) shows that Rothstein's tests perform poorly with small samples, Koedel and Betts (2011) show that his results are driven in part by non-persistent student-teacher sorting, and Chetty, Friedman and Rockoff (2014a) show that the issues he raises become less significant if "leave-year-out" measures of value-added are estimated.[11]

Studies by Kane and Staiger (2008) and Kane et al. (2013) approach the bias question from a different direction, using experiments where students are randomly assigned to teachers. Both

[11] Chetty, Friedman and Rockoff (2014a) report that in personal correspondence, Rothstein indicated that his findings are "neither necessary nor sufficient for there to be bias in a VA estimate." See Goldhaber and Chaplin (2015) and Guarino et al. (2015) for detailed analyses on this point.

studies take the same general approach. First, the authors estimate teacher effectiveness in a pre-period using standard VAMs in a non-experimental setting (their VAMs are structured like the two-step model shown in equations 4 and 5). Next, the non-experimental value-added estimates are used to predict the test scores of students who are randomly assigned to teachers in the subsequent year.[12] The 2008 study uses teacher value-added as the sole "pre-period" performance measure. The 2013 study uses a composite effectiveness measure that includes value-added and other non-test-based measures of teacher performance, although the weights on the components of the composite measure are determined based on how well they predict prior value-added and thus, value-added is heavily weighted.[13]

Under the null hypothesis of zero bias, the coefficient on the non-experimental value-added estimate in the predictive regression of student test scores under random assignment should be equal to one. Kane and Staiger (2008) estimate coefficients on the non-experimental value-added estimates in math and reading of 0.85 and 0.99, respectively, for their models that are closest to what we show in equations (4) and (5). Neither coefficient can be statistically distinguished from one. When the model is specified as a gain-score model it performs marginally worse, when it includes school fixed effects it performs marginally better in math but worse in reading, and when it is specified in test score levels with student fixed effects it performs poorly. Kane et al. (2013) obtain analogous estimates in math and reading of 1.04 and 0.70, respectively, and again, neither estimate can be statistically distinguished from one. The 2013 study does not consider a gain-score model or models that include school or student fixed effects.

While the findings from both experimental studies are consistent with the scope for bias in standard value-added estimates being small, several caveats are in order. First, sample-size and compliance issues are such that the results are noisy and the authors cannot rule out fairly large

[12] Rosters of students were randomly assigned to teachers within pairs or blocks of teachers.
[13] A case where prior value-added is the only pre-period performance measure is also considered in Kane et al. (2013).

biases at standard confidence levels.[14] Second, due to feasibility issues, the random-assignment experiments were performed within schools and among pairs of teachers for whom their principal was agreeable to randomly assigning students between them. This raises concerns about externalizing the findings to a more inclusive model to evaluate teachers (Paufler and Amrein-Beardsley, 2014; Rothstein, 2010; Rothstein and Mathis, 2013), and in particular, about the extent to which the experiments are informative about sorting across schools. Third, although these studies indicate that teacher value-added is an accurate predictor of student achievement in an experimental setting on average, this does not preclude individual prediction errors for some teachers.[15]

Chetty, Friedman and Rockoff (2014a) also examine the extent to which estimates of teacher value-added are biased. Their primary VAM is a variant of the model shown in equation (2), but they also consider a two-step VAM.[16] They do not examine models with school or student fixed effects.

Chetty, Friedman and Rockoff (2014a) take two different approaches in their investigation of the bias question. First, they merge a 21-year administrative data panel for students and teachers in a large urban school district – one that contains information similar to datasets used to estimate value-added in other locales – with tax data from the Internal Revenue Service (IRS). The tax data contain information about students and their families at a level of detail never before seen in the value-added literature. They use these data to determine the scope for omitted-variables-bias in standard models that do not include such detailed information. The key variables that Chetty,

[14] Rothstein (2010) raises this concern about Kane and Staiger (2008). Kane et al. (2013) perform a much larger experiment. While this improves statistical precision, it is still the case that large biases cannot be ruled out with 95-percent confidence in the larger study (Rothstein and Mathis, 2013).

[15] Building on this last point, in addition to general imprecision, one potential source of individual prediction errors in these studies is "overcorrection bias," which the experimental tests are not designed to detect owing to their use of two-step VAMs to produce the pre-period value-added measures. The non-experimental predictions can still be correct on average in the presence of overcorrection bias, but inaccurate for some teachers (for more information, see footnote 35 on page 33 of Kane et al., 2013). This is less of an issue for the Chetty, Friedman and Rockoff (2014a) study that we discuss next.

[16] There are some procedural differences between how Chetty, Friedman and Rockoff (2014a) estimate their models and how the models are shown in equations (2), (4) and (5). The two most interesting differences are: (1) the use of a "leave-year-out" procedure, in which a teacher's value-added in any given classroom is estimated from students in other classrooms taught by the same teacher, and (2) their allowance for teacher value-added to shift over time, rather than stay fixed in all years.

Friedman and Rockoff (2014a) incorporate into their VAMs from the tax data are mother's age at the student's birth, indicators for parental 401(k) contributions and home ownership, and an indicator for parental marital status interacted with a quartic in household income. They estimate that the magnitude of the bias in value-added estimates in typical circumstances when these parental characteristics are unobserved is 0.2 percent, with an upper-bound at the edge of the 95-percent confidence interval of 0.25 percent. The authors offer two reasons for why their estimates of the bias are so small. First, the standard controls that they use in their "district-data-only" VAM (e.g., lagged test scores, poverty status, etc.) capture much of the variation in the additional tax-data variables. Put differently, the marginal value of these additional variables is small, which suggests that the standard controls in typically-estimated VAMs are quite good. Despite this fact, however, the parental characteristics are still important independent predictors of student achievement. The second reason that the bias is small is that the variation in parental characteristics from the tax data that is otherwise unaccounted for in the model is essentially uncorrelated with teacher value-added. Chetty, Friedman and Rockoff (2014a) conclude that student-teacher sorting based on parental/family characteristics that are not otherwise accounted for in typically-estimated VAMs is limited in practice.

Chetty, Friedman and Rockoff's (2014a) second approach to examining the scope for bias in teacher value-added has similarities to the above-described experimental studies by Kane and Staiger (2008) and Kane et al. (2013), but is based on a quasi-experimental design. In instances where a staffing change occurs in a school-by-grade-by-year cell, the authors calculate the expected change in average value-added in the cell corresponding to the change. They estimate average value-added for teachers in each cell using data in all years except the years in-between which the staffing changes occur. Next, they use their out-of-sample estimate of the change in average teacher value-added at the school-by-grade level to predict changes in student achievement across cohorts of students. Put

differently, they compare achievement for different cohorts of students who pass through the same school-by-grade, but differ in the average value-added of the teachers to which they are exposed due to staffing changes.[17]

Their quasi-experimental approach requires stronger identifying assumptions than the experimental studies, as detailed in their paper. However, the authors can examine the scope for bias in value-added using a much larger dataset, which allows for substantial gains in precision. Furthermore, because the staffing changes they consider include cross-school moves, their results are informative about sorting more broadly. Using their second method, Chetty, Friedman and Rockoff (2014a) estimate that the forecasting bias from using their observational value-added measures to predict student achievement is 2.6 percent and not statistically significant (the upper bound of the 95-percent confidence interval on the bias is less than 10 percent).

Results similar to those from Chetty, Friedman and Rockoff (2014a) are reported in a recent study by Bacher-Hicks, Kane and Staiger (2014) using data from Los Angeles. Bacher-Hicks, Kane and Staiger (2014) use the same quasi-experimental switching strategy and also find no evidence of bias in estimates of teacher value-added. At the time of our writing this review, there is an ongoing debate between Bacher-Hicks, Kane and Staiger (2014), Chetty, Friedman and Rockoff (2014c) and Rothstein (2014) regarding the usefulness of the quasi-experimental teacher-switching approach for identifying teacher value-added. Rothstein (2014) argues that a correlation between teacher switching and students' prior grade test scores invalidates the approach. Chetty, Friedman and Rockoff (2014c) and Bacher-Hicks, Kane and Staiger (2014) acknowledge the correlation documented by Rothstein (2014), but find that it is mechanical and driven by the fact that prior data were used to estimate value-added. They argue that Rothstein's failure to account for the mechanical correlation makes his proposed placebo test based on prior test scores uninformative about the validity of the teacher-

[17] This approach is similar in spirit to Koedel (2008), where identifying variation in the quality of high school math teachers is caused by changes in school staffing and course assignments across cohorts.

switching approach, and present additional specifications which show how the correlation of value-added with changes in prior scores can be eliminated without substantively changing the estimated effect of value-added on changes in current scores.[18,19]

Future research will undoubtedly continue to inform our understanding of the bias issue. To date, the studies that have used the strongest research designs provide compelling evidence that estimates of teacher value-added from standard models are not meaningfully biased by student-teacher sorting along observed or unobserved dimensions.[20] It is notable that there is not any direct counterevidence indicating that value-added estimates are substantially biased.

As the use of value-added spreads to more research and policy settings, we stress that given the observational nature of the value-added approach, the absence of bias in current research settings does not preclude bias elsewhere or in the future. Bias could emerge in VAMs estimated in other settings if unobserved sorting occurs in a fundamentally different way, and the application of VAMs to inform teacher evaluations could alter sorting and other behavior (e.g., see Barlevy and

[18] For example, a teacher who moves forward from grade-4 to grade-5 within a school will have an impact on changes in current and prior scores of 5th graders; when Chetty, Friedman and Rockoff (2014c) drop the small number of teachers with these movements, this greatly reduces "placebo effects" on lagged scores but leaves effects on current scores largely unchanged. Chetty, Friedman and Rockoff (2014c) conclude that after accounting for these types of mechanical correlations, Rothstein's findings are consistent with their interpretation, pointing to results in Rothstein's appendix to support this claim.

[19] Rothstein also raises a concern about how Chetty, Friedman and Rockoff (2014a) deal with missing data. He presents an argument for why their quasi-experimental test should be conducted with value-added of zero imputed for teachers with missing estimates, rather than dropping these teachers and their students, with the former approach resulting in a larger estimate of bias. Chetty, Friedman and Rockoff (2014c) address this point by demonstrating that, because value-added is positively correlated within school-grade cells, imputing the population mean (i.e., zero) results in attenuation of the coefficient on value-added in the quasi-experimental regression. They argue that the best way to assess the importance of missing data is to examine the significant subset of school-grade cells where there is little or no missing data. This estimate of bias is quite close to zero, both in their data and in Rothstein's replication study. Bacher-Hicks, Kane and Staiger (2014) also discuss the merits of the alternative approaches to dealing with missing data in this context. In their study, "reasonably imputing the expected impacts of [teachers] with missing value-added does not change the finding regarding the predictive validity of value-added" (p. 4).

[20] In addition to the studies covered in detail in this section, which in our view are the most narrowly-targeted on the question of bias in estimates of teacher value-added, other notable studies that provide evidence consistent with value-added measures containing useful information about teacher productivity include Glazerman et al. (2013), Harris and Sass (2014), Jackson (2014), Jacob and Lefgren (2008), and Kane and Staiger (2012). Relatedly, Deming (2014) tests for bias in estimates of *school value-added* using random school choice lotteries and fails to reject the hypothesis that school effects are unbiased.

Neal, 2012; Campbell, 1976).[21] Although the quasi-experimental method developed by Chetty, Friedman and Rockoff (2014a) is subject to ongoing scholarly debate as noted above, it is a promising tool for measuring bias in VAMs without the need for experimental manipulation of classroom assignments.

3.2 Stability

Although some of the highest-profile studies on teacher value-added have focused on the issue of bias, the extent to which value-added measures will be useful in research and policy applications also depends critically on their stability. Without some meaningful degree of persistence over time, even unbiased estimates offer little value. If we assert that value added models can be developed and tested in which there is a minimal role for bias (based on the evidence discussed in the previous section), estimated teacher value-added can be described as consisting of three components: (1) real, persistent teacher quality; (2) real, non-persistent teacher quality, and (3) estimation error (which is, of course, not persistent). Of primary interest in many research studies is the real, persistent component of teacher quality, with the second and third factors contributing to the instability with which this primary parameter is estimated.[22] A number of studies have provided evidence on the stability of estimated teacher value-added over time and across schools and classrooms (e.g., see Aaronson, Barrow and Sander, 2007; Chetty, Friedman and Rockoff, 2014a; Glazerman et al., 2013; Goldhaber and Hansen, 2013; Jackson, 2014; Koedel, Leatherman and Parsons, 2012; McCaffrey et al., 2009; Schochet and Chiang, 2013).

[21] This line of criticism is not specific to value-added. The use of other high-stakes evaluation metrics could also alter behavior.

[22] Standard data and methods are not sufficient to distinguish between factors (2) and (3). For example, suppose that a non-persistent positive shock to estimated teacher value-added is observed in a given year. This shock could be driven by a real positive shock in teacher performance (e.g., perhaps the teacher bonded with her students more so than in other years) or measurement error (e.g., the lagged scores of students in the teacher's class were affected by a random negative shock in the prior year, which manifests itself in the form of larger than usual test-score gains in the current year). We would need better data than are typically available to sort out these alternative explanations. For example, data from multiple tests of the same material given on different days could be used to separate out test measurement error from teacher performance that is not persistent over classrooms/time.

Studies that document the year-to-year correlation in estimated teacher value-added have produced estimates that range from 0.18 to 0.64. Differences in the correlations across studies are driven by several factors. One factor that influences the precision with which teacher value-added can be estimated, and thus the correlation of value-added over time, is the student-teacher ratio. Goldhaber and Hansen (2013) and McCaffrey et al. (2009) document the improved precision in value-added estimates that comes with increasing teacher-level sample sizes. Goldhaber and Hansen (2013) show that the predictive value of past value-added over future value-added improves with additional years of data. The improvement is non-linear and the authors find that adding more years of data beyond three years is of limited practical value for predicting future teacher performance. Beyond the diminishing marginal returns of additional information, another explanation for this result is that as the time horizon gets longer, "drift" in real teacher performance (Chetty, Friedman and Rockoff, 2014a) puts downward pressure on the predictive power of older value-added measures. This drift will offset the benefits associated with using additional data unless the older data are properly down-weighted. McCaffrey et al. (2009) examine the benefits of additional data purely in terms of teacher-level sample sizes, which are positively but imperfectly correlated with years of data. Unsurprisingly, they document reductions in the average standard error for teacher value-added estimates as sample sizes increase.

A second factor that affects the stability of value-added estimates is whether the model includes fixed effects for students and/or schools. Adding these layers of fixed effects narrows the identifying variation used to estimate teacher value-added, which can increase imprecision in estimation. In models with student fixed effects, estimates of teacher value-added are identified by comparing teachers who share students; in models with school fixed effects, the identifying variation is restricted to occur only within schools.

Despite relying entirely on within-unit variation (i.e., school or student) for identification, school and student fixed effects models can facilitate comparisons across all teachers as long as teachers are linked across units. For example, with multiple years of data, a model with school fixed effects can be used to rank all teachers based on value-added as long as enough teachers switch schools (Mansfield, forthcoming). Teachers who share the same school are directly linked within schools, and school switchers link groups of teachers across schools. Non-switchers can be compared via "indirect linkages" facilitated by the switchers. However, the reliance on school switching to link the data comes with its own challenges and costs in terms of precision. In the context of a model designed to estimate value-added for teacher preparation programs, Mihaly et al. (2013b) note "indirect linkages can make estimates imprecise, with the potential for significant variance inflation" (p. 462).[23]

Goldhaber, Walch and Gabele (2013), Koedel, Leatherman and Parsons (2012) and McCaffrey et al. (2009) estimate the year-to-year correlation in teacher value-added from lagged-score models similar in structure to the one shown in equation (2), and from additional specifications that include school or student fixed effects. Correlations of adjacent year value-added measures estimated from models without school or student fixed effects range between 0.47 and 0.64 across these studies. McCaffrey et al. (2009) report an analogous correlation for a model that includes student fixed effects of 0.29 (using a gainscore specification).[24] Goldhaber, Walch and Gabele (2013) and Koedel, Leatherman and Parsons (2012) report correlations from models that include school fixed effects ranging from 0.18 to 0.33. Thus, while there is a wide range of stability estimates throughout the literature, much of the variance in the estimated correlations is driven by

[23] Also note that the use of student fixed effects strains the data in that it requires estimating an additional n parameters using a large-N, small-T dataset, which is costly in terms of degrees of freedom (Ishii and Rivkin, 2009).
[24] For the McCaffrey et al. (2009) study we report simple averages of the correlations shown in Table 2 across districts and grade spans.

modeling decisions. Put differently, among similarly-specified models, there is fairly consistent evidence with regard to the stability of value-added estimates.

It is also important to recognize that for many policy decisions, the year-to-year stability in value-added is not the appropriate measure by which to judge its usefulness (Staiger and Kane, 2014). As an example consider a retention decision to be informed by value-added. The relevant question to ask about the usefulness of value-added is how well a currently-available measure predicts *career* performance. How well a currently-available measure predicts next year's performance is not as important. Staiger and Kane (2014) show that year-to-year correlations in value-added are significantly lower than year-to-career correlations, and it is these latter correlations that are most relevant for judging the usefulness of value-added in many policy applications.

The question of how much instability in VAMs is "acceptable" is difficult to answer in the abstract. On the one hand, the level of stability in teacher value-added is similar to the level of stability in performance metrics widely used in other professions that have been studied by researchers, including salespeople, securities analysts, sewing-machine operators and baseball players (McCaffrey et al., 2009; Glazerman et al., 2010). However, on the other hand, any decisions based on VAM estimates will be less than perfectly correlated with the decisions one would make if value-added was known with certainty, and it is theoretically unclear whether using imperfect data in teaching is less beneficial (or more costly) than in other professions. The most compelling evidence on this question comes from studies that, taking the instability into account, evaluate the merits of acting on estimates of teacher value-added to improve workforce quality (Chetty, Friedman and Rockoff, 2014b; Boyd et al., 2011, Goldhaber and Theobald, 2013; Winters and Cowen, 2013; Rothstein, 2015). These studies consistently show that using information about teacher value-added improves student achievement relative to the alternative of not using value-added information.

4. Evidence-Based Model Selection and the Estimation of Teacher Value-Added

Section 2 raised a number of conceptual and econometric issues associated with the choice of a value-added model. Absent direct empirical evidence, a number of different specifications are defensible. This section uses the evidence presented in Section 3, along with evidence from elsewhere in the literature, to highlight areas of consensus and disagreement on key VAM specification issues.

4.1 Gain-Score Versus Lagged-Score VAMs

Gain-score VAMs have been used in a number of studies and offer some conceptual appeal (Meghir and Rivkin, 2010). One benefit is that they avoid the complications that arise from including lagged achievement, which is measured with error, as an independent variable. However, available evidence suggests that the gain-score specification does not perform as well as the lagged-score specification under a broad range of estimation conditions (Andrabi et al., 2011; Guarino, Reckase and Wooldridge, 2015; Kane and Staiger, 2008; McCaffrey et al., 2009). Put differently, the restriction on the lagged-score coefficient appears to do more harm than good. Thus, most research studies and policy applications of value-added use a lagged-score specification.[25]

It is noteworthy that the lagged-score VAM has outperformed the gain-score VAM despite the fact that most studies do not implement any procedures to directly account for measurement error in the lagged-score control. The measurement-error problem is complicated because test measurement error in standard test instruments is heteroskedastic (Andrabi et al., 2011; Boyd et al., 2013; Lockwood and McCaffrey, 2014; Koedel, Leatherman and Parsons, 2012). Lockwood and McCaffrey (2014) provide the most comprehensive investigation of the measurement-error issue in the value-added context of which we are aware. Their study offers a number of suggestions for ways

[25] In addition to the fact that most research studies cited in this review use a lagged-score VAM, lagged-score VAMs are also the norm in major policy applications where VAMs are incorporated into teacher evaluations – e.g., see the VAMs that are being estimated in Washington DC (Isenberg and Walsh, 2014), New York City (Value Added Research Center, 2010) and Pittsburgh (Johnson et al., 2012).

to improve model performance by directly accounting for measurement error that should be given careful consideration in future VAM applications. One straightforward finding in Lockwood and McCaffrey (2014) is that controlling for multiple prior test scores can mitigate the influence of test measurement error. A simple but important takeaway from Koedel, Leatherman and Parsons (2012) is that one should not assume that that measurement error in the lagged test score is constant across the test-score distribution. Although this assumption simplifies the measurement-error correction procedure, it is not supported by the data and can worsen model performance.

4.2 Student and School Fixed Effects

Chetty, Friedman and Rockoff (2014a) show that a VAM along the lines of the one shown in equation (2) – without school and student fixed effects – produces estimates of teacher value-added with very little bias (forecasting bias of approximately 2.6 percent, which is not statistically significant). Kane et al. (2013) and Kane and Staiger (2008) also show that models without school and student fixed effects perform well. While the experimental studies are not designed to assess the value of using school fixed effects because the randomization occurs within schools, they provide evidence consistent with unbiased estimation within schools without the need to control for unobserved student heterogeneity using student fixed effects.

Thus, although school and student fixed effects are conceptually appealing because of their potential to reduce bias in estimates of teacher value-added, the best evidence to date suggests that their practical importance is limited. Without evidence showing that these layers of fixed effects reduce bias it is difficult to make a compelling case for their inclusion in VAMs. Minus this benefit, it seems unnecessary to bear the cost of reduced stability in teacher value-added associated with including them, per the discussion in Section 3.2.

4.3 One-Step Versus Two-Step VAMs

Models akin to the standard, one-step VAM in equation (2) and the two-step VAM in equations (4) and (5) perform well in the validity studies by Chetty, Friedman and Rockoff (2014a), Kane and Staiger (2008) and Kane at al. (2013). Chetty, Friedman and Rockoff (2014a) is the only study of the three to examine the two modeling structures side-by-side – they show that value-added estimates from both models exhibit very little bias. Specifically, their estimates of forecasting bias are 2.6 and 2.2 percent for the one-step and two-step VAMs, respectively, and neither estimate is statistically significant. If the primary objective is bias reduction, available evidence does not point to a clearly-preferred modeling structure.

Although we are not aware of any studies that directly compare the stability of value-added estimates from one-step and two-step models, based on the discussion in Section 3.2 we do not expect significant differences to emerge along this dimension either. The reason is that the two fundamental determinants of stability that have been established by the literature – teacher-level sample sizes and the use of school and/or student fixed effects – will not vary between the one-step and two-step VAM structures.

Thus, the literature does not indicate a clearly preferred modeling choice along this dimension. This is driven in large part by the fact that the two modeling structures produce similar results – Chetty, Friedman and Rockoff (2014a) report a correlation in estimated teacher value-added across models of 0.98 in their data.[26] For most research applications, the high correlation offers sufficient grounds to support either approach. However, in policy applications this may not be the case, as even high correlations allow for some (small) fraction of teachers to have ratings that are substantively affected by the choice of modeling structure (Castellano and Ho, 2015).

[26] This result is based on a comparison of fully-specified versions of the two types of models. Larger differences may emerge when using sparser specifications.

4.4 Covariates

We briefly introduced the conditioning variables that have been used by researchers in VAM specifications with equation (2). Of those variables, available evidence suggests that the most important, by far, are the controls for prior achievement (Chetty, Friedman and Rockoff, 2014a; Ehlert et al., 2013; Lockwood and McCaffrey, 2014; Rothstein, 2009).

Rothstein (2009) and Ehlert et al. (2013) show that multiple years of lagged achievement in the same subject are significant predictors of current achievement conditional on single-lagged achievement in the same subject, which indicates that model performance is improved by including these variables. Ehlert et al. (2013) note that "moving from a specification with three lagged scores to one with a single lagged score does not systematically benefit or harm certain types of schools or teachers (p. 19)," but this result is sensitive to the specification that they use. Lockwood and McCaffrey (2014) consider the inclusion of a more comprehensive set of prior achievement measures from multiple years and multiple subjects. They find that including the additional prior-achievement measures attenuates the correlation between estimated teacher value-added and the average prior same-subject achievement of teachers' students.

Chetty, Friedman and Rockoff (2014a) examine the sensitivity of value-added to the inclusion of different sets of lagged-achievement information but use data only from the prior year. Specifically, rather than extending the time horizon to obtain additional lagged-achievement measures, they use lagged achievement in another subject along with aggregates of lagged achievement at the school and classroom levels. Similarly to Ehlert et al. (2013), Lockwood and McCaffrey (2014) and Rothstein (2009), they show that the additional measures of prior achievement are substantively important in their models. In one set of results they compare VAMs that include each student's own lagged achievement in two subjects, with and without school- and classroom-average lagged achievement in these subjects, to a model that includes only the student's

own lagged achievement in the same subject. The models do not include any other control variables. Their quasi-experimental estimate of bias rises from 3.82 percent (statistically insignificant) with all controls for prior achievement, to 4.83 percent (statistically insignificant) with only controls for a student's own prior scores in both subjects, to 10.25 percent (statistically significant) with only controls for a student's prior score in the same subject. Notably, their estimate of bias in the model that includes the detailed prior-achievement measures, but no other covariates, is not far from their estimate based on the full model (2.58 percent, statistically insignificant).

Compared to the lagged-achievement controls, whether to include the additional demographic and socioeconomic controls that are typically found in VAMs is a decision that has been shown to be of limited practical consequence, at least in terms of global inference (Aaronson, Barrow and Sander, 2007; Ehlert et al., 2013; McCaffrey et al., 2004). However, two caveats to this general result are in order. First, the degree to which including demographic and socioeconomic controls in VAMs is important will depend on the richness of the prior-achievement controls in the model. Models with less information about prior achievement may benefit more from the inclusion of demographic and socioeconomic controls. Second, and re-iterating a point from above, in policy applications it may be desirable to include demographic and socioeconomic controls in VAMs, despite their limited impact on the whole, in order to guard against teachers in the most disparate circumstances being systematically identified as over- or under-performing. For personnel evaluations, it seems statistically prudent to use all available information about students and the schooling environment to mitigate as much bias in value-added estimates as possible. This is in contrast to federal guidelines for state departments of education issued in 2009, which discourage

the use of some control variables in value-added models (United States Department of Education, 2009).[27]

5. What we Know About Teacher Value-Added

5.1 Teacher Value-Added is an Educationally and Economically Meaningful Measure

A consistent finding across research studies is that there is substantial variation in teacher performance as measured by value-added. Much of the variation occurs within schools (e.g., see Aaronson, Barrow and Sander, 2007; Rivkin, Hanushek and Kain, 2005). Hanushek and Rivkin (2010) report estimates of the standard deviation of teacher value-added in math and reading (when available) across ten studies and conclude that the literature leaves "little doubt that there are significant differences in teacher effectiveness" (p. 269). The effect of a one-standard deviation improvement in teacher value-added on student test scores is estimated to be larger than the effect of a ten student reduction in class size (Rivkin, Hanushek and Kain, 2005; Jepsen and Rivkin, 2009).[28]

Research has demonstrated that standardized test scores are closely related to school attainment, earnings, and economic outcomes (e.g., Chetty et al., 2011; Hanushek and Woessmann, 2008; Lazear, 2003; Murnane et al., 2000), which supports the test-based measurement of teacher effectiveness. Hanushek (2011) combines the evidence on the labor market returns to higher cognitive skills with evidence on the variation in teacher value-added to estimate the economic value of teacher quality. With a class size of 25 students, he estimates the economic value of a teacher who

[27] The federal guidelines are likely politically motivated, at least in part – see Ehlert et al. (forthcoming). Also note that some potential control variables reflect student designations that are partly at the discretion of education personnel. An example is special-education status. The decision of whether to include "discretionary" variables in high-stakes VAMs requires accounting for the incentives to label students that can be created by their inclusion.

[28] Although the immediate effect on achievement of exposure to a high-value-added teacher is large, it fades out over time. A number of studies have documented rapid fade-out of teacher effects after one or two years (Kane and Staiger, 2008; Jacob, Lefgren and Sims, 2010; Rothstein, 2010). Chetty, Friedman and Rockoff (2014b) show that teachers' impacts on test scores stabilize at approximately 25 percent of the initial impact after 3-4 years. It is not uncommon for the effects of educational interventions to fade out over time (e.g., see Currie and Thomas, 2000; Deming, 2009; Krueger and Whitmore, 2001).

is one-standard deviation above the mean to exceed $300,000 across a range of parameterizations that allow for sensitivity to different levels of true variation in teacher quality, the decay of teacher effects over time, and the labor-market returns to improved test-score performance.

Chetty, Friedman and Rockoff (2014b) complement Hanushek's work by directly estimating the long-term effects of teacher value-added. Following on their initial validity study (Chetty, Friedman and Rockoff, 2014a), they use tax records from the Internal Revenue Service (IRS) to examine the impact of exposure to teacher value-added in grades 4-8 on a host of longer-term outcomes. They find that a one-standard deviation improvement in teacher value-added in a single grade raises the probability of college attendance at age 20 by 2.2 percent and annual earnings at age-28 by 1.3 percent. For an entire classroom, their earnings estimate implies that replacing a teacher whose performance is in the bottom 5 percent in value-added with an average teacher would increase the present discounted value of students' lifetime earnings by $250,000.[29] In addition to looking at earnings and college attendance, they also show that exposure to more-effective teachers reduces the probability of teenage childbearing, increases the quality of the neighborhood in which a student lives as an adult, and raises participation in 401(k) savings plans.

5.2 Teacher Value-Added is Positively but Imperfectly Correlated Across Subjects and Across Different Tests Within the Same Subject

Goldhaber, Cowan and Walch (2013) correlate estimates of teacher value-added across subjects for elementary teachers in North Carolina. They report that the within-year correlation between math and reading value-added is 0.60. After adjusting for estimation error, which attenuates the correlational estimate, it rises to 0.80. Corcoran, Jennings and Beveridge (2011) and Lefgren and Sims (2012) also show that value-added in math and reading are highly correlated. The latter study

[29] The Chetty, Friedman and Rockoff (2014b) estimate is smaller than the comparable range of estimates from Hanushek (2011). Two key reasons are (1) Hanushek uses a lower discount rate for future student earnings, and (2) Hanushek allows for more persistence in teacher impacts over time. Nonetheless, both studies indicate that high-value-added teachers are of substantial economic value.

leverages the cross-subject correlation to determine an optimal weighting structure for value-added in math and reading given the objective of predicting future value-added for teachers most accurately.

Teacher value-added is also positively but imperfectly correlated across different tests within the same subject. Lockwood et al. (2007) compare teacher value-added on the "procedures" and "problem solving" subcomponents of the same mathematics test.[30] They report correlations of value-added estimates across a variety of different VAM specifications. Estimates from their standard one-step, lagged-score VAM indicate a correlation of teacher value-added across testing subcomponents of approximately 0.30.[31] This correlation is likely to be attenuated by two factors. First, as noted by the authors, there is a lower correlation between student scores on the two subcomponents of the test in their analytic sample relative to the national norming sample as reported by the test publisher. Second, the correlational estimate is based on estimates of teacher value-added that are not adjusted for estimation error, an issue that is compounded by the fact that the subcomponents of the test have lower reliability than the full test.[32]

Papay (2011) replicates the Lockwood et al. (2007) analysis with a larger sample of teachers and shrunken value-added estimates. The cross-subcomponent correlation of teacher value-added in his study is 0.55. Papay (2011) reports that most of the differences in estimated teacher value-added across test subcomponents cannot be explained by differences in test content or scaling, and thus

[30] The authors note that "procedures items cover computation using symbolic notation, rounding, computation in context and thinking skills, whereas problem solving covers a broad range of more complex skills and knowledge in the areas of measurement, estimation, problem solving strategies, number systems, patterns and functions, algebra, statistics, probability, and geometry. The two sets of items are administered in separately timed sections." (p. 49)

[31] This is an average of their single-year estimates from the covariate adjusted model with demographics and lagged test scores from Table 2.

[32] Lockwood et al. (2007) report that the subcomponent test reliabilities are high, but not as high as the reliability for the full test.

concludes that the primary factor driving down the correlation is test measurement error.[33] Corcoran, Jennings and Beveridge (2011) perform a similar analysis and obtain similar results using student achievement on two separate math and reading tests administered in the Houston Independent School District. They report cross-test, same-subject correlations of teacher value-added of 0.59 and 0.50 in math and reading, respectively. The tests they evaluate also differ in their stakes, and the authors show that the estimated variance of teacher value-added is higher on the high-stakes assessment in both subjects.

The practical implications of the correlations in teacher value-added across subjects and instruments are not obvious. The fact that the correlations across all alternative achievement metrics are clearly positive is consistent with there being an underlying generalizable teacher-quality component embodied in the measures. This is in line with what one would expect given the influence of teacher value-added on longer-term student outcomes (Chetty, Friedman and Rockoff, 2014b). However, it is disconcerting that different testing instruments can lead to different sets of teachers being identified as high- and low-performers, as indicated by the Corcoran, Jennings and Beveridge (2011), Lockwood et al. (2007) and Papay (2011) studies. Papay's (2011) conclusion that the lack of consistency in estimated value-added across assessments is largely the product of test measurement error seems reasonable, particularly when one considers the full scope for measurement error in student test scores (Boyd et al., 2013).

5.3 Teacher Value-Added is Positively but Imperfectly Correlated with Other Evaluation Metrics

Several studies have examined correlations between estimates of teacher value-added and alternative evaluation metrics. Harris and Sass (2014) and Jacob and Lefgren (2008) correlate teacher value-added with principal ratings from surveys. Jacob and Lefgren (2008) report a correlation of

[33] Papay (2011) also compares value-added measures across tests that are given at different points in the school year and shows that test timing is an additional factor that contributes to differences in teacher value-added as estimated across testing instruments.

0.32 between estimates of teacher value-added in math and principal ratings based on principals' beliefs about the ability of teachers to raise math achievement; the analogous correlation for reading is 0.29 (correlations are reported after adjusting for estimation error in the value-added measures). Harris and Sass (2014) report slightly larger correlations for the same principal assessment – 0.41 for math and 0.44 for reading – and also correlate value-added with "overall" principal ratings and report correlations in math and reading of 0.34 and 0.38, respectively.[34]

Data from the Measures of Effective Teaching (MET) project funded by the Bill and Melinda Gates Foundation have been used to examine correlations between teacher value-added and alternative evaluation metrics including classroom observations and student surveys. Using data primarily from middle school teachers who teach multiple sections within the same year, Kane and Staiger (2012) report that the correlation between teacher scores from classroom observations and value-added in math, measured across different classrooms, ranges from 0.16 to 0.26 depending on the observation rubric. The correlation between value-added in math and student surveys of teacher practice, also taken from different classrooms, is somewhat larger at 0.37. Correlations between value-added in reading and these alternative metrics remain positive but are lower. They range from 0.10 to 0.25 for classroom observations across rubrics, and the correlation between student survey results and reading value-added is 0.25.[35]

Kane et al. (2011) provide related evidence by estimating value-added models to identify the effects of teaching practice on student achievement. Teaching practice is measured in their study by the observational assessment used in the Cincinnati Public Schools' Teacher Evaluation System (TES). Cincinnati's TES is one of the few examples of a long-standing high stakes assessment with

[34] The question from the Harris and Sass (2014) study for the overall rating makes no mention of test scores: "Please rate each teacher on a scale from 1 to 9 with 1 being not effective to 9 being exceptional" (p. 188).
[35] Polikoff (2014) further documents that the correlations between these alternative evaluation measures and value-added vary across states, implying that different state tests are differentially sensitive to instructional quality as measured by these metrics.

external peer evaluators, and thus it is quite relevant for public policy (also see Taylor and Tyler, 2012). Consistent with the evidence from the MET project, the authors find that observational measures of teacher practice are positively correlated with value-added. Their preferred models that use a composite index to measure classroom practice imply that a 2.3 standard-deviation increase in the index (one point) corresponds to a little more than a one-standard-deviation increase in teacher value-added in reading, and a little less than a one-standard-deviation increase in math.[36]

The positive correlations between value-added and the alternative, non-test-based metrics lends credence to the informational value of these metrics in light of the strong validity evidence emerging from the research literature on value-added. Indeed, out-of-sample estimates of teacher quality based on principal ratings, observational rubrics and student surveys positively predict teacher impacts on student achievement, although not as well as out-of-sample value-added (Harris and Sass, 2014; Jacob and Lefgren, 2008; Kane et al., 2011; Kane and Staiger, 2012). At present, no direct evidence on the longer-term effects of exposure to high-quality teachers along these alternative dimensions is available.[37]

6. Policy Applications

6.1 Using Value-Added to Inform Personnel Decisions in K-12 Schools

The wide variation in teacher value-added that has been consistently documented in the literature has spurred research examining the potential benefits to students of using information about value-added to improve workforce quality. The above-referenced studies by Hanushek (2011) and Chetty, Friedman and Rockoff (2014b) perform direct calculations of the benefits associated with removing ineffective teachers and replacing them with average teachers (also see Hanushek,

[36] Rockoff and Speroni (2011) also provide evidence that teachers who receive better subjective evaluations produce greater gains in student achievement. A notable aspect of their study is that some of the evaluation metrics they consider come from prior to teachers being hired into the workforce.

[37] Because many of these metrics are just emerging in rigorous form (e.g., see Weisberg et al., 2009), a study comparable to something along the lines of Chetty et al. (2014b) may not be possible for some time. In the interim, non-cognitive outcomes such as persistence in school and college-preparatory behaviors, in addition to cognitive outcomes, can be used to vet these alternative measures similarly to the recent literature on VAMs.

2009). Their analyses imply that using estimates of teacher-value added to inform personnel decisions would lead to substantial gains in the production of human capital in K-12 schools. Similarly, Winters and Cowen (2013) use simulations to show how incorporating information about teacher value-added into personnel policies can lead to improvements in workforce quality over time.

Monetizing the benefits of VAM-based removal policies and comparing them to costs suggests that they will easily pass a cost-benefit test (Chetty, Friedman and Rockoff, 2014b).[38] A caveat is that the long-term labor supply response is unknown. Rothstein (2015) develops a structural model to describe the labor market for teachers that incorporates costs borne by workers from the uncertainty associated with teacher evaluations based on value-added. He estimates that teacher salaries would need to rise under a more-rigorous, VAM-based personnel policy. The salary increase would raise costs, but not by enough to offset the benefits of the improvements in workforce quality (Chetty, Friedman and Rockoff, 2014b; Rothstein, 2015). However, as noted by Rothstein (2015), his findings are highly sensitive to how he parameterizes his model. This limits inference given the general lack of evidence in the literature with regard to several of his key parameters – in particular, the elasticity of labor supply and the degree of foreknowledge that prospective teachers possess about their own quality. As more-rigorous teacher evaluation systems begin to come online, monitoring the labor-supply response will be an important area of research. Initial evidence from the most mature high-stakes system that incorporates teacher value-added into personnel decisions – the IMPACT evaluation system in Washington, DC – provides no indication of an adverse labor-supply response of yet (Dee and Wyckoff, 2013).

[38] Studies by Boyd et al. (2011) and Goldhaber and Theobald (2013) consider the use of value-added measures to inform personnel policies within the more narrow circumstance of forced teacher layoffs. Both studies show that layoff policies based on value-added result in significantly higher student achievement relative to layoff policies based on seniority. Seniority-driven layoff policies were a central point of contention in the recent Vergara v. California court case.

In addition to developing rigorous teacher evaluation systems that formally incorporate teacher value-added, there are several other ways that information about value-added can be leveraged to improve workforce quality. For example, Rockoff et al. (2012) show that simply providing information to principals about value-added increases turnover for teachers with low performance estimates and decreases turnover for teachers with high estimates. Correspondingly, average teacher productivity increases. Student achievement increases in line with what one would expect given the workforce-quality improvement.

Condie, Lefgren and Sims (2014) suggest an alternative use of value-added to improve student achievement. Although they perform simulations indicating that VAM-based removal policies will be effective (corroborating earlier studies), they also argue that value-added can be leveraged to improve workforce quality without removing any teachers. Specifically, they propose to use value-added to match teachers to students and/or subjects according to their comparative advantages (also see Goldhaber, Cowan and Walch, 2013). Their analysis suggests that data driven workforce re-shuffling has the potential to improve student achievement by more than a removal policy targeted at the bottom 10-percent of teachers.

Glazerman et al. (2013) consider the benefits of a different type of re-shuffling based on value-added. They study a program that provides pecuniary bonuses to high-value-added teachers who are willing to transfer to high-poverty schools from other schools. The objective of the program is purely equity-based. Glazerman et al. (2013) show that the teacher transfer policy raises student achievement in high-poverty classrooms that are randomly assigned to receive a high-value-added transfer teacher.[39]

[39] There is some heterogeneity underlying this result. Specifically, Glazerman et al. (2013) find large, positive effects for elementary classrooms that receive a transfer teacher but no effects for middle-school classrooms.

6.2 Teacher Value-Added as a Component of Combined Measures of Teaching Effectiveness

Emerging teacher evaluation systems in practice are using value-added as one component of a larger "combined measure." Combined measures of teaching effectiveness also typically include non-test-based performance measures like classroom observations, student surveys and measures of professionalism.[40] Many of the non-test-based measures are newly emerging, at least in rigorous form (see Weisberg et al., 2009). The rationale for incorporating multiple measures into teacher evaluations is that teaching effectiveness is multi-dimensional and no single measure can capture all aspects of quality that are important.

Mihaly et al. (2013a) consider the factors that go into constructing an optimal combined measure based on data from the MET project. The components of the combined measure that they consider are teacher value-added on several assessments, student surveys and observational ratings. Their analysis highlights the challenges associated with properly weighting the various components in a combined-measure evaluation system. Assigning more weight to any individual component makes the combined measure a stronger predictor of that component in the future, but a weaker predictor of the other components. Thus, the optimal approach to weighting depends in large part on value judgments about the different components by policymakers.[41]

The most mature combined-measure evaluation system in the United States, and the one that has received the most attention nationally, is the IMPACT evaluation system in Washington, DC. IMPACT evaluates teachers based on value-added, teacher-assessed student achievement, classroom observations and a measure of commitment to the school community.[42] Teachers

[40] School districts that are using or in the process of developing combined measures of teaching effectiveness include Los Angeles (Strunk, Weinstein, and Makkonen, 2014), Pittsburgh (Scott and Correnti, 2013) and Washington DC (Dee and Wyckoff, 2013); state education agencies include Delaware, New Mexico, Ohio, and Tennessee (White, 2014).
[41] Polikoff (2014) further considers the extent to which differences in the sensitivity of state tests to the quality and/or content of teachers' instruction should influence how weights are determined in different states. He recommends that states explore their own data to determine the sensitivity of their tests to high-quality instruction.
[42] The weight on value-added in the combined measure was 35 percent for teachers with value-added scores during the 2013-2014 school year, down from 50 percent in previous years (as reported by Dee and Wyckoff, 2013). The 15 percent

evaluated as ineffective under IMPACT are dismissed immediately, teachers who are evaluated as minimally effective face pressure to improve under threat of dismissal (dismissal occurs with two consecutive minimally-effective ratings), and teachers rated as highly effective receive a bonus the first time and an escalating bonus for two consecutive highly-effective ratings (Dee and Wyckoff, 2013). Using a regression discontinuity design, Dee and Wyckoff (2013) show that teachers who are identified as minimally effective a single time are much more likely to voluntarily exit, and for those who do stay, they are more likely to perform better in the following year.[43] Dee and Wyckoff (2013) also show that teachers who receive a first highly-effective rating improve their performance in the following year, a result that the authors attribute to the escalating bonus corresponding to two consecutive highly-effective ratings.

In Washington DC, at least during the initial years of IMPACT, there is no *prima facie* evidence of an adverse labor-supply effect. Dee and Wyckoff (2013) report that new teachers in Washington DC entering after the second year of IMPACT's implementation were rated as significantly more effective than teachers who left (about one-half of a standard deviation of IMPACT scores). The labor-supply response in Washington DC could be due to the simultaneous introduction of performance bonuses in IMPACT and/or idiosyncratic aspects of the local teacher labor market. Thus, while this is an interesting early result, labor-supply issues bear careful monitoring as similar teacher evaluation systems emerge in other locales across the United States.

7. Conclusion

This article has reviewed the literature on teacher value-added, covering issues ranging from technical aspects of model design to the use of value-added in public policy. A goal of the review has been to highlight areas of consensus and disagreement in research. Although a broad spectrum of

---

drop in the weight on value-added for eligible teachers was offset in 2013-2014 by a 15-percent weight on teacher-assessed student achievement, which was not used to evaluate value-added eligible teachers in prior years.

[43] Dee and Wyckoff's research design is not suited to speak to whether the improvement represents a causal effect on performance or is the result of selective retention.

views is reflected in available studies, along a number of important dimensions the literature appears to be converging on a widely-accepted set of facts.

Perhaps the most important result for which consistent evidence has emerged in research is that students in K-12 schools stand to gain substantially from policies that incorporate information about value-added into personnel decisions for teachers (Boyd et al., 2011; Chetty, Friedman and Rockoff, 2014b; Condie, Lefgren and Sims, 2014; Dee and Wyckoff, 2013; Glazerman, 2013; Goldhaber, Cowan and Walch, 2013; Goldhaber and Theobald, 2013; Hanushek, 2009, 2011; Rothstein, 2015; Winters and Cowen, 2013). The consistency of this result across the variety of policy applications that have been considered in the literature is striking. In some applications the costs associated with incorporating value-added into personnel policies would likely be small – examples include situations where mandatory layoffs are required (Boyd et al., 2011; Goldhaber and Theobald, 2013), and the use of value-added to adjust teaching responsibilities rather than make what are likely to be more-costly and less-reversible retention/removal decisions (Condie, Lefgren and Sims, 2014; Goldhaber, Cowan and Walch, 2013). However, even in more costly applications, such as the use of value-added to inform decisions about retention and removal, available evidence suggests that the benefits to students of using value-added to inform decision-making will outweigh the costs (Chetty, Friedman and Rockoff, 2014b; Rothstein, 2015).

In addition to the growing consensus among researchers on this critical issue, our review also uncovers a number of other areas where the literature is moving toward general agreement (many of which serve as the building blocks for the consensus described in the previous paragraph). For example, the research studies that have employed the strongest experimental and quasi-experimental designs to date indicate that the scope for bias in estimates of teacher value-added

from standard models is quite small.[44] Similarly, aggregating available evidence on the stability of value-added reveals consistency in the literature along this dimension as well. In addition, agreement among researchers on a number of model-specification issues has emerged, ranging from which covariates are the most important in the models to whether gain-score or lagged-score VAMs produce more reliable estimates.

Still, our understanding of a number of issues would benefit from more evidence. An obvious area for expansion of research is into higher grades, as the overwhelming majority of studies focus on students and teachers in elementary and middle schools. The stronger sorting that occurs as students move into higher grades will make estimating value-added for teachers in these grades more challenging, but not impossible. There are also relatively few studies that examine the portability of teacher value-added across schools (e.g., see Jackson, 2013; Xu, Ozek and Corritore, 2012). A richer evidence base on cross-school portability would be helpful for designing policies aimed at, among other things, improving equity in access to effective teachers (as in Glazerman et al., 2013). Another area worthy of additional exploration is the potential to improve instructional quality by using information about value-added to inform teacher assignments, ranging from which subjects to teach to how many students to teach (Condie, Lefgren and Sims, 2014; Goldhaber, Cowan and Walch, 2013; Hansen, 2014). More work is also needed on the relationships between value-added and alternative evaluation metrics, like those used in combined measures of teaching effectiveness (e.g., classroom observations, student surveys, etc.). Using value-added to validate these measures, and determining how to best combine other measures with value-added to evaluate teachers, can help to inform decision-making in states and school districts working to develop and implement more rigorous teacher evaluation systems.

[44] As described above, the caveat to this result is that the absence of bias in current research settings does not preclude bias elsewhere or in the future. Nonetheless, available evidence to date offers optimism about the ability of value-added to capture performance differences across teachers.

As state and district evaluation systems come online and start to mature, it will be important to examine how students, teachers and schools are affected along a number of dimensions (with Dee and Wyckoff, 2013, serving as an early example). Heterogeneity in implementation across systems can be leveraged to learn about the relative merits of different types of personnel policies, ranging from retention/removal policies to merit pay to the data driven re-shuffling of teacher responsibilities. Studying these systems will also shed light on some of the labor-market issues raised by Rothstein (2015), about which we currently know very little.

In summary, the literature on teacher value-added has provided a number of valuable insights for researchers and policymakers, but much more remains to be learned. Given what we already know, this much is clear: the implementation of personnel policies designed to improve teacher quality is a high-stakes proposition for students in K-12 schools. It is the students who stand to gain the most from efficacious policies, and who will lose the most from policies that come up short.

References

Aaronson, Daniel, Lisa Barrow and William Sander. 2007. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25(1), 95-135.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja and Tristan Zajonc. 2011. Do Value-Added Estimates Add Value? Accounting for Learning Dynamics. *American Economic Journal: Applied Economics* 3(3): 29-54.

Bacher-Hicks, Andrew, Thomas J. Kane and Douglas O. Staiger. 2014. Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles. NBER Working Paper No. 20657.

Baker, Eva L., Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard. 2010. Problems with the Use of Student Test Scores to Evaluate Teachers. Economic Policy Institute Briefing Paper No. 278.

Barlevy, Gary and Derek Neal. 2012. Pay for Percentile. *American Economic Review* 102(5), 1805-31.

Ben-Porath, Yoram. 1967. The Production of Human Capital and the Life-Cycle of Earnings. *Journal of Political Economy* 75(4), 352-365.

Betts, Julian R. and Y. Emily Tang. Value-Added and Experimental Studies of the Effect of Charter Schools on Student Achievement: A Literature Review. 2008. Bothell, WA: National Charter School Research Project, Center on Reinventing Public Education.

Betts, Julian R., Andrew C. Zau and Kevin King. 2005. *From Blueprint to Reality: San Diego's Education Reforms*. San Francisco, CA: Public Policy Institute of California.

Biancarosa, Gina, Anthony S. Byrk and Emily Dexter. 2010. Assessing the Value-Added Effects of Literacy Collaborative Professional Development on Student Learning. *Elementary School Journal* 111(1), 7-34.

Boyd, Donald, Hamilton Lankford, Susanna Loeb and James Wyckoff. 2011. Teacher Layoffs: An Empirical Illustration of Seniority v. Measures of Effectiveness. *Education Finance and Policy* 6(3), 439-454.

Boyd, Donald, Hamilton Lankford, Susanna Loeb and James Wyckoff. 2013. Measuring Test Measurement Error: A General Approach. *Journal of Educational and Behavioral Statistics* 38(6), 629-663.

Campbell, Donald T. 1976. Assessing the Impact of Planned Social Change. Occasional Paper Series #8). Hanover, NH: The Public Affairs Center, Dartmouth College.

Castellano, Katherine E. and Andrew D. Ho. 2015. Practical Differences Among Aggregate-Level Conditional Status Metrics: From Median Student Growth Percentiles to Value-Added Models. *Journal of Educational and Behavioral Statistics* 40(1), 35-68.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanual Saez, Diane Whitmore Schanzenbach and

Danny Yagan. 2011. How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *Quarterly Journal of Economics* 126(4), 1593-1660.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2014a. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review* 104(9), 2593-2632.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2014b. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review* 104(9), 2633-79.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2014c. Response to Rothstein (2014) on 'Revisiting The Impacts of Teachers.' Unpublished manuscript, Harvard University.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor. 2006. Teacher-Student Matching and the Assessment of Teacher Effectiveness. *Journal of Human Resources* 41(4): 778-820.

Condie, Scott, Lars Lefgren and David Sims. 2014. Teacher Heterogeneity, Value-Added and Education Policy. *Economics of Education Review* 40(1), 76-92.

Corcoran, Sean, Jennifer L. Jennings and Andrew A. Beveridge. 2011. Teacher Effectiveness on High- and Low-Stakes Tests. Unpublished Manuscript, New York University.

Currie, Janet and Duncan Thomas. 2000. School Quality and the Longer-Term Effects of Head Start. *Journal of Human Resources* 35(4), 755-774.

Dee, Thomas and James Wyckoff. 2013. Incentives, Selection and Teacher Performance. Evidence from IMPACT. NBER Working Paper No. 19529.

Deming, David J. 2009. Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics* 1(3), 111-134.

Deming, David J. 2014. Using School Choice Lotteries to Test Measures of School Effectiveness. *American Economic Review* 104(5): 406-411.

Ehlert, Mark, Cory Koedel, Eric Parsons and Michael Podgursky (forthcoming). Selecting Growth Measures for Use in School Evaluation Systems: Should Proportionality Matter? *Educational Policy*.

Ehlert, Mark, Cory Koedel, Eric Parsons and Michael Podgursky. 2014. Choosing the Right Growth Measure: Methods Should Compare Similar Schools and Teachers. *Education Next* 14(2), 66-71.

Ehlert, Mark, Cory Koedel, Eric Parsons and Michael Podgursky. 2013. The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence from School- and Teacher-Level Models in Missouri. *Statistics and Public Policy* 1(1): 19-27.

Glazerman, Steven, Susanna Loeb, Dan Goldhaber, Douglas Staiger, Stephen Raudenbush, and Grover Whitehurst. 2010. Evaluating Teachers: The Important Role of Value-Added. Policy Report. Brown Center on Education Policy at Brookings.

Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, Jeffrey Max and Elizabeth Warner. 2013. Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. United States Department of Education.

Goldhaber, Dan and Duncan Chaplin. 2015. Assessing the "Rothstein Falsification Test." Does it Really Show Teacher Value-added Models are Biased? *Journal of Research on Educational Effectiveness* 8(1), 8-34.

Goldhaber, Dan, James Cowan and Joe Walch. 2013. Is a Good Elementary Teacher Always Good? Assessing Teacher Performance Estimates Across Subjects. *Economics of Education Review* 36(1), 216-228.

Goldhaber, Dan and Michael Hansen. 2013. Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. *Economica* 80(319), 589-612.

Goldhaber, Dan, Stephanie Liddle and Roddy Theobald. 2013. The Gateway to the Profession: Evaluating Teacher Preparation Programs Based on Student Achievement. *Economics of Education Review* 34(1), 29-44.

Goldhaber, Dan and Roddy Theobald. 2013. Managing the Teacher Workforce in Austere Times: The Determinants and Implications of Teacher Layoffs. *Education Finance and Policy* 8(4), 494–527.

Goldhaber, Dan, Joe Walch and Brian Gabele. 2013. Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments. *Statistics and Public Policy* 1(1): 28-39.

Guarino, Cassandra M., Mark D. Reckase, Brian W. Stacey, and Jeffrey W. Wooldridge. 2015. Evaluating Specification Tests in the Context of Value-Added Estimation. *Journal of Research on Educational Effectiveness* 8(1), 35-59.

Guarino, Cassandra M., Mark D. Reckase and Jeffrey W. Wooldridge. 2015. Can Value-Added Measures of Teacher Performance be Trusted? *Education Finance and Policy* 10(1), 117-156.

Guarino, Cassandra, Michelle Maxfield, Mark D. Reckase, Paul Thompson and Jeffrey M. Wooldridge. 2014. An Evaluation of Empirical Bayes' Estimation of Value-Added Teacher Performance Measures. Unpublished manuscript.

Hansen, Michael. 2014. Right-Sizing the Classroom: Making the Most of Great Teachers. CALDER Working Paper No. 110.

Hanushek, Eric A. 2011. The Economic Value of Higher Teacher Quality. *Economics of Education Review* 30(3), 266-479.

Hanushek, Eric A. 2009. Teacher Deselection, in *Creating a New Teaching Profession* eds. Dan Goldhaber and Jane Hannaway. Urban Institute, Washington, DC.

Hanushek, Eric A. 1979. Conceptual and Empirical Issues in the Estimation of Educational Production Functions. *The Journal of Human Resources* 14(3), 351-388.

Hanushek, Eric A. and Steven G. Rivkin. 2010. Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review* 100(2), 267-271.

Hanushek, Eric A. and Ludger Woessmann. 2008. The Role of Cognitive Skills in Economic Development. *Journal of Economic Literature* 46(3), 607-668.

Harris, Douglas N. and Tim R. Sass. 2011. Teacher Training, Teacher Quality, and Student Achievement. *Journal of Public Economics* 95: 798-812.

Harris, Douglas N. and Tim R. Sass. 2014. Skills, Productivity and the Evaluation of Teacher Performance. *Economics of Education Review* 40, 183-204.

Herrmann, Mariesa, Elias Walsh, Eric Isenberg and Alexandra Resch. 2013. Shrinkage of Value-Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels. Policy Report, Mathematica Policy Research.

Isenberg, Eric and Elias Walsh. 2014. Measuring School and Teacher Value Added in DC, 2012-2013 School Year: Final Report. Policy Report, Mathematica Policy Research (01.17.2014).

Ishii, Jun and Steven G. Rivkin. 2009. Impediments to the Estimation of Teacher Value Added. *Education Finance and Policy* 4(4), 520-536.

Jackson, Kirabo. 2013. Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers. *Review of Economics and Statistics* 95(4), 1096-1116.

Jackson, Kirabo. 2014. Teacher Quality at the High-School Level: The Importance of Accounting for Tracks. *Journal of Labor Economics* 23(4), 645-684.

Jacob, Brian A. and Lars Lefgren. 2008. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluations in Education. *Journal of Labor Economics* 26(1), 101-136.

Jacob, Brian A., Lars Lefgren and David P. Sims. 2010. The Persistence of Teacher-Induced Learning Gains. *Journal of Human Resources* 45(4), 915-943.

Jepsen, Christopher and Steven G. Rivkin. 2009. Class Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size. *Journal of Human Resources* 44(1), 223-250.

Johnson, Matthew, Stephen Lipscomb, Brian Gill, Kevin Booker and Julie Bruch. 2012. Value-Added Model for Pittsburgh Public Schools. Unpublished report, Mathematica Policy Research.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller and Douglas O. Staiger. 2013. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, Tom J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. What Does Certification Tell us about Teacher Effectiveness? Evidence from New York City. *Economics of Education Review* 27(6), 615-631.

Kane, Thomas J. and Douglas O. Staiger. 2012. Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, Thomas J. and Douglas O. Staiger. 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER Working Paper No. 14607.

Kane, Thomas J. and Douglas O. Staiger. 2002. The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives* 16(4), 91-114.

Kane, Thomas J., Eric S. Taylor, John H. Tyler and Amy L. Wooten. 2011. Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources* 46(3), 587-613.

Kinsler, Joshua. 2012. Assessing Rothstein's Critique of Teacher Value-Added Models. *Quantitative Economics* 3(2), 333-362.

Koedel, Cory. 2009. An Empirical Analysis of Teacher Spillover Effects in Secondary School. *Economics of Education Review* 28(6), 682-692.

Koedel, Cory. 2008. Teacher Quality and Dropout Outcomes in a Large, Urban School District. *Journal of Urban Economics* 64(3), 560-572.

Koedel, Cory and Julian R. Betts. 2011. Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy* 6(1), 18-42.

Koedel, Cory, Eric Parsons, Michael Podgursky and Mark Ehlert (forthcoming). Teacher Preparation Programs and Teacher Quality: Are There Real Differences Across Programs? *Education Finance and Policy*.

Koedel, Cory, Rebecca Leatherman and Eric Parsons. 2012. Test Measurement Error and Inference from Value-Added Models. *B.E. Journal of Economic Analysis & Policy*, 12(1) (Topics).

Konstantopoulos, S., & Chung, V. 2011. The Persistence of Teacher Effects in Elementary Grades. *American Educational Research Journal* 48, 361–386

Krueger, Alan B. and Diane M. Whitmore. 2001. The Effect of Attending a Small Class in the Early Grades on College Test Taking and Middle School Test Results: Evidence from Project STAR. *Economic Journal* 111(468), 1-28.

Lazear, E. P. (2003). Teacher incentives. *Swedish Economic Policy Review* 10(3), 179–214.

Lefgren, Lars and David P. Sims. 2012. Using Subject Test Scores Efficiently to Predict Teacher Value-Added. *Educational Evaluation and Policy Analysis* 34(1), 109-121.

Lockwood, J.R. and Daniel F. McCaffrey. 2014. Correcting for Test Score Measurement Error in ANCOVA Models for Estimating Treatment Effects. *Journal of Educational and Behavioral Statistics* 39(1), 22-52.

Lockwood, J.R., Daniel F. McCaffrey, Laura S. Hamilton, Brian Stecher, Vi-Nhuan Le and Jose Felipe Martinez. 2007. The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement* 44(1), 47-67.

Mansfield, Richard K. (forthcoming). Teacher Quality and Student Inequality. *Journal of Labor Economics*.

McCaffrey, Daniel F., Lockwood, J. R., Koretz, Daniel M., Louis, Thomas A., & Hamilton, Laura 2004. Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics* 29(1), 67–101.

McCaffrey, Daniel F., Tim R. Sass, J.R. Lockwood and Kata Mihaly. 2009. The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy* 4(4), 572-606.

Meghir, Costas and Steven G. Rivkin. 2011. Econometric Methods for Research in Education in *Handbook of the Economics of Education* (volume 3) eds. Eric A. Hanushek, Stephen Machin and Ludger Woessmann. Waltham, MA: Elsevier.

Mihaly, Kata, Daniel F. McCaffrey, Douglas O. Staiger and J.R. Lockwood. 2013a. A Composite Estimator of Effective Teaching. *RAND External Publication*, EP-50155.

Mihaly, Kata, Daniel McCaffrey, Tim R. Sass and J.R. Lockwood. 2013b. Where You Come From or Where You Go? Distinguishing Between School Quality and the Effectiveness of Teacher Preparation Program Graduates. *Education Finance and Policy* 8(4), 459-493.

Mihaly, Kata, Daniel F. McCaffrey, Tim R. Sass, and J.R. Lockwood. 2010. Centering and Reference Groups for Estimates of Fixed Effects: Modifications to felsdvreg. *The Stata Journal* 10(1): 82-103.

Morris, Carl N. 1983. Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association* 78(381), 47–55.

Murnane, Richard J., John B. Willett, Yves Duhaldeborde and John H. Tyler. 2000. How important are the cognitive skills of teenagers in predicting subsequent earnings*? Journal of Policy Analysis and Management* 19(4), 547–568.

Newton, Xiaoxia A., Linda Darling-Hammond, Edward Haertel and Ewart Thomas. 2010. Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability Across Models and Contexts. *Education Policy Analysis Archives* 18(23), 1-27.

Nye, Barbara, Spyros Konstantopoulos and Larry V. Hedges. 2004. How Large are Teacher Effects? *Educational Evaluation and Policy Analysis* 26(3), 237-257.

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

Paufler, Noelle A. and Audrey Amrein-Beardsley. 2014. The Random Assignment of Students into Elementary Classrooms: Implications for Value-Added Analyses and Interpretations. *American Educational Research Journal* 51(2), 328-362.

Polikoff, Morgan S. 2014. Does the Test Matter? Evaluating Teachers When Tests Differ in Their Sensitivity to Instruction. In T. J. Kane, K. A. Kerr, & R. C.Pianta (Eds.). *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project* (pp. 278 - 302). San Francisco, CA: Jossey-Bass.

Rivkin, Steven G., Eric A. Hanushek and John F. Kain. 2005. Teachers, Schools and Academic Achievement. *Econometrica* 73(2), 417-58.

Rockoff, Jonah E. and Cecilia Speroni. 2011. Subjective and Objective Evaluations of Teacher Effectiveness: Evidence from New York City. *Labour Economics* 18(5), 687-696.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane and Eric S. Taylor. 2012. Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. *American Economic Review* 102(7), 3184-3213.

Rothstein, Jesse. 2015. Teacher Quality Policy When Supply Matters. *American Economic Review* 105(1), 100-130.

Rothstein, Jesse. 2014. Revising the Impacts of Teachers. Unpublished manuscript, University of California-Berkeley.

Rothstein, Jesse. 2010. Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics* 125(1), 175-214.

Rothstein, Jesse. 2009. Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy* 4(4), 537-571.

Rothstein, Jesse and William J. Mathis. 2013. Review of Two Culminating Reports from the MET Project. National Education Policy Center Report.

Sass, Tim R., Jane Hannaway, Zeyu Xu, David N. Figlio and Li Feng. 2012. Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools. *Journal of Urban Economics* 72, 104-122.

Sass, Tim R., Anastasia Semykina and Douglas N. Harris. 2014. Value-Added Models and the Measurement of Teacher Productivity. *Economics of Education Review* 38(1), 9-23.

Schochet, Peter Z. and Hanley S. Chiang. 2013. What are Error Rates for Classifying Teacher and School Performance Measures Using Value-Added Models? *Journal of Educational and Behavioral Statistics* 38(3), 142-171.

Scott, Amy and Richard Correnti. 2013. Pittsburgh's New Teacher Improvement System: Helping Teachers Help Students Learn. Policy Report. Pittsburgh, PA: A+ Schools.

Staiger, Douglas O. and Thomas J. Kane. 2014. Making Decisions with Imprecise Performance Measures: The Relationship Between Annual Student Achievement Gains and a Teacher's Career Value-Added. In Thomas J. Kane, Kerri A. Kerr and Robert C. Pianta (Eds.) *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project* (p. 144-169). San Francisco, CA: Jossey-Bass.

Strunk, Katharine O., Tracey L. Weinstein and Reino Makkonnen. 2014. Sorting out the Signal: Do Multiple Measures of Teachers' Effectiveness Provide Consistent Information to Teachers and Principals? *Education Policy Analysis Archives* 22(100).

Taylor, Eric S. and John H. Tyler. 2012. The Effect of Evaluation on Teacher Performance. *American Economic Review* 102(7), 3628-3651.

Todd, Petra E. and Kenneth I. Wolpin. 2003. On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal* 113, F3-F33.

United States Department of Education. 2009. Growth Models: Non-Regulatory Guidance. Unpublished.

Value Added Research Center. 2010. NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model. Unpublished report, Wisconsin Center for Education Research, University of Wisconsin-Madison.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern and David Keeling. 2009. The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. New York: The New Teacher Project.

White, Taylor. 2014. Evaluating Teachers More Strategically. Using Performance Results to Streamline Evaluation Systems. Policy Report. Carnegie Foundation for the Advancement of Teaching.

Winters, Marcus A. and Joshua M. Cowen. 2013. Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies. *Journal of Policy Analysis and Management* 32(3), 634-654.

Xu, Zeyu, Umut Ozek and Matthew Corritore. 2012. Portability of Teacher Effectiveness Across School Settings. CALDER Working Paper No. 77.