Where is the Value in Value-Added Modeling?

White Paper

Dan Murphy

April 2012

About Pearson

Pearson, the global leader in education and education technology, provides print and digital education materials for preK through college, student information systems and learning management systems, teacher licensure testing, teacher professional development, career certification programs, and testing and assessment products that set the standard for the industry. Pearson's other primary businesses include the Financial Times Group and the Penguin Group. For more information about the Assessment & Information group of Pearson, visit http://www.pearsonassessments.com

About Pearson's White Papers

Pearson's white paper series shares the perspectives of our assessment experts on selected topics of interest to educators, researchers, poilcymakers and other stakeholders involved in assessment and instruction. Pearson's publications in .pdf format may be obtained at: http://www.pearsonassessments.com/research

Abstract

There is currently tremendous interest in developing value-added models that use student scores on standardized tests to estimate the effects that teachers have on student learning. However, little consensus exists within the research community about which value-added models are most appropriate or the manner in which they should be incorporated into accountability systems. In short, recent policies encouraging the use of these models to evaluate teachers have been controversial. This paper first provides an overview of value-added modeling (VAM), including definitions and descriptions of three general types of value-added models. The paper then summarizes the rationales for and against incorporating VAM into an accountability framework. Finally, we provide a recommendation that VAM estimates be used in conjunction with other measures to form a composite. Policymakers considering the use of VAM in their accountability systems should view VAM results as formative tools that help teachers identify areas of strength and weakness to assist in professional development.

Keywords: value-added modeling, teacher effectiveness, accountability

Where is the Value in Value-Added Modeling?

One of the greatest shifts in educational reform over the last decade has been the move from compliance-based accountability systems, under which educators are expected to comply with governmental rules and regulations, toward results-based accountability systems. Under results-based accountability systems, teachers are accountable for student learning and accountable to the general public (Anderson, 2005). The shift toward results-based accountability systems is based on the idea that objective measurement of student performance is the best way to measure the performance of teachers and schools, and that associating consequences with student performance outcomes motivates better performance. There is tremendous interest in developing statistical models that use student scores on standardized tests to estimate the effects that teachers have on student learning. The overarching term that describes this statistical approach is value-added modeling (VAM).

VAM uses the results from repeated measurements of student performance to estimate the effects of individual teachers on students' learning trajectories. The models are designed to answer the question of whether the influence of a particular school or teacher causes students' achievement to grow more or less than the "average" or "expected" growth. McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004) provide a description of a general value-added model, presented below. The expected score *y* for student *i* in grade *g* can be expressed as:

$$E(y_{ig}) = \boldsymbol{\mu}_g + \boldsymbol{\beta}'_g \mathbf{x}_i + \boldsymbol{\gamma}'_{gg} \mathbf{z}_{ig}$$

where $\mathbf{x}_i$ are time invariant covariates such as gender and ethnicity, $\mathbf{z}_{ig}$ are time varying covariates such as economic disadvantage and prior test scores, and $\boldsymbol{\mu}_g$ is the average score in grade g after accounting for the time invariant and time varying covariates. In other words, Equation 1 states that an individual student's expected score in grade g is dependent only on the time invariant and time varying covariates, $\mathbf{x}_i$ and $\mathbf{z}_{ig}$.

The student's achieved score in any particular grade, however, can be expressed as a function of the effects of his or her particular school and teacher in addition to the time invariant and time varying covariates, as presented in Equation 2:

$$y_{ig} = \boldsymbol{\mu}_g + \boldsymbol{\beta}'_g \mathbf{x}_i + \boldsymbol{\gamma}'_{gg} \mathbf{z}_{ig} + \boldsymbol{\lambda}'_g \boldsymbol{\eta}_g + \boldsymbol{\varphi}'_{ig} \boldsymbol{\theta}_g + \boldsymbol{\varepsilon}_{ig}$$

In Equation 2, $\boldsymbol{\eta}_g$ represents the effects of the student's school, and $\boldsymbol{\theta}_g$ represents the effects of the student's teacher. The term $\boldsymbol{\varepsilon}_{ig}$ represents random error and is assumed to be normally distributed with a mean of zero. The fact that the student's expected score in Equation 1 is dependent only on the values of his or her covariates implies that the school effect, $\boldsymbol{\eta}_g$, and teacher effect, $\boldsymbol{\theta}_g$, in Equation 2 represent "deflections" from the student's expected growth trajectory. These deflections are inferred to be the value that the school and teacher add to the student's growth trajectory (Raudenbush and Bryk, 2002). The deflections can be positive or negative. The values added by the school and teacher are inferred to be measures of their causal effect on a student's achievement. Stated slightly differently, VAM estimates the causal effects that schools and teachers have by measuring deflections from the growth trajectory the student was

expected to have under the average teacher (McCaffrey, Lockwood, Koretz, and Hamilton, 2003).

There are many different statistical models that can be considered to be special cases of the general value-added model presented in Equation 2. The models that are most commonly discussed within the research literature can be categorized under three broad classifications: gain score models, residualized gain score models, and multivariate longitudinal models. There are advantages and disadvantages associated with each model type, and none of the models has been singled out for superior performance across all conditions. Table 1 presents the defining features, advantages and disadvantages, and prototype examples of the three model types.

Table 1

Value-Added Model Classifications and Examples with Associated Advantages and Disadvantages

Model Type	Defining Features	Advantages	Disadvantages	Examples
Gain Score	• Test score gains from one year to the next are compared • Gains are compared across teachers • Assumes all scores are on the same scale • The test score data are cross-sectional as opposed to longitudinal	• Intuitive to stakeholders • Student growth is the variable of interest • Simple models that are easy to implement and understand • Accounts for factors that affect only students' level of achievement	• Ignores important information about students • Test scores must be vertically scaled • Students with missing data are excluded from the model	• Simple Gain Score Models • Cross-Sectional Gain Score
Residualized Gain Score	• Current scores are regressed on prior score and the residuals are considered to be deflections from expected growth • Often incorporate student characteristics into the model • The test score data are cross-sectional as opposed to longitudinal	• Can be thought of as predicting achievement with the average teacher even if scores are not on the same scale • Simple to specify and fit • Can be specified as non-liner	• Fitting the models separately for each year of data ignores important information, e.g., past teacher influence • Students who are missing data are excluded from the model	• Covariate Adjustment Models • Dallas Value-Added Accountability System
Multiariate Longitudinal Models	• Model the full joint distribution of student outcomes • The test score data are cross-sectional as opposed to longitudinal	• Can handle missing data • Use all available information about students • Conservative teacher effects estimates are intended to reduce teacher misclassifications	• If data are not missing at random, the results will be biased • Extreme computational burden • Only special cases can be fit using standard mixed model software and even then only to modestly-sized data sets	• Education Value-Added Assessment System • Cross-Classified Random Effects Models

The use of these models to hold teachers accountable for student performance has not been without controversy. In short, there is currently little consensus regarding the ability of VAM to accurately and reliably capture the effects of individual teachers. In the sections that follow we discuss the rationales for and against the use of VAM to measure teacher effects and provide recommendations to policymakers who may be considering using VAM to evaluate teachers. We describe arguments from both sides within this controversy. First, we illustrate how VAM is superior to the results-based accountability system known as attainment-based accountability and the teacher effectiveness measures that are currently in place. This is followed by research finding that teacher effects are significant and may impact student success beyond a single year (Chetty, Friedman, and Rockoff, 2011; Sanders and Rivers, 1996; Mendro, Jordan, Gomez, Anderson, and Bembry, 1998). Next, arguments against the use of VAM are described, including the points that students are not randomly assigned to classrooms and schools, correct model specification is always in doubt, and high-stakes accountability decisions based on VAM may have unintended consequences (Braun, 2009; Feng, Figlio, and Sass, 2010; Finnigan and Gross, 2007; Kupermintz, 2003).

Reasons to Use VAM to Evaluate Teachers

VAM Is Superior to Attainment-Based Accountability Systems

Results-based accountability can be broken down into two sub-categories: attainment-based and growth-based accountability systems. VAM's increasing popularity is due in part to the clearer picture of student learning (i.e., growth) it portrays as a growth-based system when compared with attainment-based accountability systems such as the No Child Left Behind Act of 2001 (NCLB) and its Adequate Yearly Progress (AYP) requirements. In contrast to growth-based accountability systems that track student cohorts across time, attainment-based accountability systems such as AYP compare the test performance from the current year to the test performance from the previous year within the same grade level. A valid criticism of the attainment-based accountability approach is that it compares things that are not necessarily comparable, because student populations may be significantly different from one year to the next (Braun, 2005). For example, school residency boundary changes could result in a systematic change to the student population such that a school's average test scores increase or decrease before the school year begins. Attainment-based accountability systems would be unable to account for such a scenario and may mistakenly attribute yearly changes in test performance to teacher effectiveness that would be more properly attributed to school boundary changes.

A second criticism of attainment-based accountability is that it does not take into account that students begin each school year in different places. Evaluating a teacher's performance by examining student academic attainment levels can be misleading, because some classrooms may consist of students who enter with high levels of achievement, and others may consist of students who enter with achievement levels well behind those of their peers. Attainment-based accountability systems cannot account for situations such as a classroom of students who enter with low achievement levels and make great gains, yet still fall short of the required proficiency cut score at the end of the year. Nor can they account for classrooms of students who enter with high achievement levels and make little progress, yet still remain above the required proficiency cut

score at the end of the year. As these scenarios illustrate, attainment-based accountability systems may inadvertently punish teachers responsible for great educational outcomes and reward those responsible for poor educational outcomes (Linn, 2005).

Because attainment-based accountability systems compare different populations and do not control for prior achievement, they are a poor measure of the effect of an individual teacher's contribution to student achievement. By contrast, VAM is designed to evaluate the learning growth of a population of students, control for factors beyond the teacher's control, and isolate the contributions that teachers make toward student test score gains. As a result, VAM provides a measure that enables comparisons across different teachers, and inferences can be drawn such as teacher X stimulates more learning than teacher Y. The comparisons enabled by VAM provide a fairer and more objective measure of teacher performance than attainment-based systems. The objective nature of VAM estimates also make them preferable to traditional and more subjective measures of teacher performance such as principal evaluations, which are more prone to being influenced by factors other than teacher performance, such as the personal relationship between the principal and teacher.

VAM Is Superior to Traditional Measures of Teacher Effectiveness

Traditionally, teacher evaluation has been predominantly based on subjective ratings. The most common method of evaluating teachers is principal evaluations that are based on relatively few classroom visits and observations per year. Student achievement is not directly evaluated or considered in this process. As with all such subjective evaluations, the personal feelings and preconceived notions of the principal can interfere with the fairness and objectivity of the evaluation. Furthermore, recent research has found little differentiation among teacher ratings from evaluation systems based solely on subjective teacher evaluation. For example, a 2007 assessment of teacher evaluations in Chicago Public Schools by the New Teacher Project found that 93% of the system's teachers were rated as "excellent" or "superior," whereas less than 0.5% of teachers received an "unsatisfactory" rating. Similarly, Weisberg, Sexton, Mulhern, and Keeling (2009) found 99% of teachers to be rated satisfactory in a system with two categories, and for systems based on more than two categories, 94% of teachers were rated in one of the top two.

The homogeneity of the evaluation results combined with the so-called "Lake Wobegon effect" (Tucker, 1997), where all teachers are considered to be above average, call the validity of the process into question. Tucker exposed this lack of validity in a study that compared principals' beliefs about teacher effectiveness with the formal evaluation process. She found estimates of the number of incompetent teachers based on principals' beliefs to be much higher than the number of teachers who were formally identified as incompetent. In particular, principals reported 5% percent of teachers being incompetent but formally documented only 2.65% of teachers as being incompetent. Tucker's findings raise concerns that the status quo may be masking a problem that educationally harms untold numbers of students.

Concerns with the teacher evaluation status quo are coming to light in part because of the trend toward results-based accountability systems (Anderson, 2005), which have spurred demands to base high-stakes personnel decisions such as teacher promotion, dismissal, and salary increases on measures of teacher effectiveness. Stakeholders interested in rewarding outstanding teachers and/or removing poor teachers from the system will not find traditional forms of teacher evaluation

to be useful. Proponents of VAM argue that it provides a fair and useful means of differentiating performance among teachers, something that school systems have traditionally been reluctant, or incapable, of doing (Goldhaber, 2010).

VAM Research Indicates Teachers Have Large and Long-Lasting Effects

Advocates of VAM cite studies showing that selecting teachers on the basis of value-added measures can increase student achievement (Gordon, Kane, and Staiger, 2006; Hanushek, 2009). VAM research also indicates that the effect sizes attributed to teachers are large when compared with other factors that influence student achievement (Hill, Kapitula, and Umland, 2011; Mendro et al., 1998; Rockoff, 2004; Wright, Horn, and Sanders, 1997), and that these effects accumulate and persist into the future (Chetty et al., 2011; Sanders and Rivers, 1996; Mendro et al., 1998).

A recent example of this type of research that has attracted considerable attention is a study by Chetty et al. (2011) who found that teachers have large effects on student learning in grades 4 to 8. They also found that teachers with high value-added measures who transfer to new schools increase achievement at their new schools in accordance with predictions. Perhaps the most ground-breaking aspect of their study examined the effect that teachers with high value-added measures have on their students' outcomes in adulthood, although it should be noted that there is some skepticism within the research community regarding this claim (see, e.g., Ballou, 2012). Chetty et al. report that students assigned to highly effective teachers in their elementary or middle school years are more likely to attend college, attend higher-ranked colleges, earn higher salaries, live in higher SES neighborhoods, and save more money for retirement. They are also less likely to have children as teenagers.

Not only did Chetty et al. (2011) find that the effects of highly effective teachers persist into adulthood, they also found that replacing ineffective teachers has a similar impact. For example, they estimate that replacing a teacher whose estimated value-add is in the bottom 5% of the teacher distribution with an average teacher would result in an increase in the students' lifetime income by more than $250,000. Thus, Chetty et al. conclude that test score gains can be used to capture the causal effects of teachers, and the long-term causal effects of teachers are profound. Good teachers create substantial economic value, and VAM is helpful in identifying good teachers.

Reasons Not to Use VAM to Evaluate Teachers

Students Are Not Randomly Assigned to Teachers

Measuring teacher effectiveness under experimental conditions would be relatively straightforward. Under such conditions, students would be randomly assigned to classrooms, and then a pre-test/post-test design could be used with the teacher considered to be a treatment variable. Statistical tests regarding the significance of the variability of students' scores between teachers would then be sufficient to make valid decisions about the effects of individual teachers, because it could be assumed that other confounding variables were randomly distributed across classrooms. In other words, the assignment of students to teachers would not advantage or disadvantage any particular teacher. Unfortunately, randomly assigning students to teachers for the purpose of measuring teacher effectiveness is rarely feasible (or ethical), and students

are not randomly assigned to classrooms in practice. Therefore, teacher effectiveness must be estimated under less than ideal conditions.

Because students are not randomly assigned to teachers, student performance is correlated with classrooms (i.e., students who perform similarly tend to "cluster" together). Statistical models cannot easily separate compositional effects due to the clustering of students from teacher effects. It is possible, for example, for schools to assign students who fail state assessments to remedial classes. Conversely, schools may use specified performance levels on state assessments as prerequisites to advanced or accelerated classes. Furthermore, it has been found that successful teachers tend to be able to select their assignments to some extent (Goldhaber and Anthony, 2004) and as a result are more likely to have highly motivated students. It is difficult to determine when students are systematically sorted into classes whether higher average levels of achievement are due to teacher effectiveness or the characteristics of the students they teach. A research study by Rothstein (2009) illustrates this point.

Rothstein (2009) developed falsification tests for value-added models based on the idea that future teachers cannot influence past student learning. Using a set of North Carolina elementary school student test score data, Rothstein found that students' 5th-grade teachers were correlated with their 4th-grade test score gains. In other words, this finding suggests that a student's future teacher relates to their current performance. The VAM results indicated that students' 5th-grade teachers taught with varying degrees of effectiveness during a school year in which they did not interact with the students. Rothstein's finding strongly suggests that students are sorted into classrooms based on their previous test score gains, and that value-added models do not completely control statistically for this sorting, even after controlling for factors known to correlate with achievement. As a result, test score gains resulting from nonrandom sorting of students into classrooms may be mistakenly attributed to teacher effectiveness. There is real concern among critics of VAM that no statistical model, however complex, and no method of analysis, however sophisticated, can fully compensate for the fact that students are not randomly assigned to schools and classrooms.

Correct Model Specification Is Always in Doubt

Model specification refers to the process of converting a theory into a statistical model by selecting the appropriate functional form and the variables to include to best fit the data. If the model specification process is done incorrectly, then the model is said to be misspecified. The problem of model misspecification has two aspects. First, there are virtually an infinite number of models that can be specified to evaluate the data. For example, teacher effects can be specified as "fixed effects" or "random effects." Fixed effects are appropriate if the estimated teacher effects are specific only to the teachers under consideration, whereas random effects are appropriate if the results are going to be generalized to a larger population of teachers. In addition, teacher effects can be considered to be cumulative and persist indefinitely or to decay gradually over time. There has also been considerable debate within the research community about whether student demographic variables should be included as covariates within the model, or whether that practice encourages lower expectations for historically underserved populations (Ballou, Sanders, and Wright, 2004; Braun, 2005; McCaffrey et al., 2003; Ross, Stringfield, Sanders, and Wright, 2003; Tekwe et al., 2004). Overall there is little consensus within the field about which model specification produces the most accurate value-added measures, and it is important to note that different model specifications can lead to different teacher effectiveness

rankings, meaning model choice can directly impact teachers' ratings (Briggs and Domingue, 2011; McCaffrey et al., 2004; Tekwe et al., 2004).

A second aspect of model misspecification in VAM, as in all nonexperimental causal modeling, is the risk of biased estimates due to omitted causative variables (Hibpshman, 2004). A model that omits variables that are causative in nature is by definition misspecified. McCaffrey et al. (2003) note that omitted variables that are randomly distributed should have little effect on the results of value-added models. However, when omitted variables cluster by class, or when they differ by strata, none of the models are capable of disentangling teacher effects from the effects of student-level covariates such as student demographic and family background characteristics. While research results can describe generally what happens when causative variables are omitted from a particular model, determining whether causative variables are omitted in practice is impossible. It is always possible, regardless of the safeguards taken to specify a model correctly, that causative variables have been inadvertently omitted.

To summarize, the experts who study VAM disagree about how best to specify value-added models and whether the models are sufficiently reliable. There are numerous possible sources of bias that can emerge when models are misspecified, and correct model specification is always in doubt.

VAM May Produce Unintended Consequences

The VAM measures discussed in this paper are norm referenced rather than criterion referenced, meaning teachers are "graded on a curve." The measure for a particular teacher only indicates where that teacher falls in comparison with other teachers in the system. In other words, a teacher's measure will be influenced in large part by the company he or she keeps. For example, a mediocre teacher measured with a weak group of teachers would obtain a better value-added measure than a similarly mediocre teacher measured with a strong group of teachers. Therefore, critics of VAM contend that it is unclear that rewards or sanctions based on value-added measures would be given to the most deserving teachers. The lack of criterion-related evaluation measures was noted by Kupermintz (2003), who cautioned, "Questions about fairness and equity must be raised if personnel decisions employ normative information that imply in practice different standards or benchmarks in different school systems" (p. 290).

In addition, VAM critics are concerned that the norm-referenced nature of VAM may provide disincentives to teacher collaboration and more incentives for teacher competition. Further, if the value-added models are not capable of disentangling the effects of sorting or student background variables, as the research cited above indicates, teachers who serve the neediest students are likely to be evaluated more harshly by the system. Consequently, there is concern that the entanglement of student and teacher effects may unintentionally tempt teachers to seek the strongest students at the expense of the neediest students (Kupermintz, 2003). There is speculation that teachers will avoid difficult assignments at the neediest schools, and within schools teachers may avoid working with students who are likely to pull down their VAM-related teacher effectiveness scores (Baker et al., 2010).

Finally, VAM critics worry that teachers may become demoralized if they perceive the evaluation system to be unfair or arbitrary. Recent survey data indicate that accountability pressures are associated with higher attrition and reduced morale, especially among teachers in high-need schools (Feng et al., 2010; Finnigan and Gross, 2007). Paradoxically, rather than motivating better

teacher performance, it is possible that high-stakes accountability based on VAM could cause talented teachers to leave the profession entirely (Baker et al., 2010). As Braun (2009) succinctly put it, "we will do students and their families no favor if we impose an accountability system that unfairly penalizes schools that are contributing to student development broadly conceived, that hastens the departure of good teachers from the field and discourages prospective teachers from entering the field altogether" (p. 55).

Recommendations

The debate concerning the use of VAM to evaluate teachers is intense and contentious. There is a danger that the contentiousness of the issue could influence policy discussions toward two polarized, but mistaken, positions (Briggs and Domingue, 2011). First, those who support the use of VAM to evaluate teachers may mistakenly attribute criticism of value-added models to support of the current teacher evaluation system. Second, those who critique value-added models may argue that the models are flawed and therefore should not be incorporated into high-stakes teacher evaluation decisions. Both of these arguments are erroneous for a similar reason related to a status quo that most researchers agree is flawed: value-added estimates need not be perfect to contribute at least in part to an improved teacher evaluation system (Briggs and Domingue).

The research base is currently insufficient to support the use of VAM as the sole basis for high-stakes decisions. There are numerous possible sources of error when attributing causal effects to teachers based on a single value-added model estimate. When teachers' livelihoods are at stake, the risks of misclassification based on value-added model estimates alone appear to outweigh the objectivity and precision that make quantitative estimates so attractive. Nevertheless, it is not clear that VAM estimates would be more harmful than the status quo. Therefore, we recommend that VAM estimates be used in conjunction with other measures to form a composite. Furthermore, we recommend that VAM results be used as formative tools that help teachers identify areas of strength and weakness to assist in their professional development.

For example, VAM may be appropriate for diagnostic purposes, such as identifying teachers who might be extremely low or high performing so that follow-up evaluations can be done to verify the VAM findings. In other words, these estimates may be useful as indicators to initiate further inquiry into why a teacher's estimate was particularly high or low and may be a good starting point for administrators to target teachers for more thorough review (McCaffrey et al., 2003). At the same time, administrators should be sensitive to patterns of identification that suggest bias may be an influence. For example, if the majority of high-performing teachers teach specialized populations such as gifted and talented or accelerated students, and low-performing teachers predominantly teach specialized populations targeted for remediation or other intervention services, then administrators should be mindful of the possibility that the models are identifying student clusters rather than teacher effectiveness.

If teachers are going to be evaluated for high-stakes decisions using VAM, then we recommend following the guidelines put forth in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). According the Standards, multiple forms of supporting evidence

should be presented to demonstrate that test scores are fulfilling their intended interpretation and not resulting in unintended negative consequences. Furthermore, Phillips (2009) suggests that there are steps that can be taken to provide sufficient evidence to satisfy a court, in the event of legal challenges to high-stakes decisions based on VAM estimates. She recommends accumulating data for a teacher across multiple classes or cohorts, providing a process for reviewing the decisions that considers extenuating circumstances, and using multiple criteria as corroborating evidence. Phillips's recommendations seem to be a reasonable starting point for a defensible framework within which VAM estimates could be supported in a high-stakes environment. As Braun (2009) put it, "What is called for is a rethinking of how to best improve student outcomes through the monitoring of multiple fallible indicators while also enlisting the professionalism of school staff" (p. 55).

In summary, we recognize that VAM may be a useful tool for evaluating teacher effectiveness and can help to improve the system for evaluating teachers that is currently in place. We recommend that the weight that is placed upon VAM estimates as a measure of teacher effectiveness be appropriate to the stakes that will accompany its outcomes. In a low-stakes environment, VAM estimates may be a good starting point for identifying teachers who may be especially effective or egregiously ineffective for the purposes of identifying best practices or targeted interventions. However, as the stakes associated with VAM increase, we recommend combining VAM estimates with multiple indicators so that the weight associated with VAM estimates decreases. In addition to student achievement, other indicators that reflect the diverse types of evidence that illustrate good teaching include: performance-based observations; surveys of students, parents, and staff; portfolios; local indicators; and peer-to-peer reviews. The appropriate weighting scheme for a multiple measure teacher evaluation system is a serious issue that may best be decided by using a standard setting approach similar to the process used to set performance standards on standardized tests. For a more in-depth discussion about developing weighted composite measures of teacher effectiveness, see Evaluating Teachers and Principals: Developing Fair, Valid, and Reliable Systems (Pearson, 2012). The most defensible position in a high-stakes teacher evaluation environment would appear to be based on multiple sources of corroborating evidence, with no one piece of evidence unduly weighted.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, J. A., International, I. E. P., & International Academy of Education. (2005). *Accountability in education*. Paris: International Institute for Educational Planning.

Baker, E. L., Barton, P. E., Darling-Hammond, L.,Haertel, E., Ladd, H. F., Linn, R., . . . Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing paper 278). Washington, D.C.: Economic Policy Institute. Accessed on February 6, 2012, from http://epi.3cdn.net/b9667271ee6c154195_t9m6iij8k.pdf

Ballou, D. (2012). *Review of The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*. Boulder, CO: National Education Policy Center. Retrieved February 16, 2012, from http://nepc.colorado.edu/thinktank/review-long-term-impacts

Ballou, D., Sanders, W. L., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics,* 29(1), 37–66.

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models.* Princeton, NJ: Educational Testing Service.

Braun, H. (2009). Discussion: With choices come consequences. *Educational Measurement: Issues and Practice,* 28(4), 52–55.

Briggs, D., & Domingue, B. (2011). *Due Diligence and the Evaluation of Teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the* Los Angeles Times. Boulder, CO: National Education Policy Center. Retrieved February 2, 2012, from http://nepc.colorado.edu/publication/due-diligence

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood.* NBER Working Papers, December 2011.

Feng, L., Figlio, D., & Sass, T. (2010). *School Accountability and Teacher Mobility.* CALDER Working Paper No. 47, June. Washington DC: CALDER. Accessed on February 10, 2012, from http://www.urban.org/uploadedpdf/1001396-school-accountability.pdf

Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. *American Educational Research Journal*, 44 (3), September: 594–630.

Goldhaber, D. (2010). When the Stakes are High, Can We Rely on Value-Added? Exploring the Use of Value-Added Models to Inform Teacher Workforce Decisions. *Center for American Progress.* Accessed on January 9, 2012, from http://www.americanprogress.org/issues/2010/12/pdf/vam.pdf

Goldhaber, D., & Anthony, E. (2004). *Can teacher quality be effectively assessed? National Board certification as a signal of effective teaching.* Seattle, WA: Center on Reinventing Public Education.

Gordon, R. J., Kane, T. J., and Staiger, D. O. (2006). "Identifying Effective Teachers Using Performance on the Job." Washington: Brookings Institution.

Hanushek, E. A. (2009). Teacher Deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a New Teaching Profession* (pp. 165–180). Washington, DC: Urban Institute Press.

Hibpshman, T.L. (2004). A Review of Value-Added Models. Kentucky Education Professional Standards Board.

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

Kupermintz, H. (2003). Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis,* 25, 287–298.

Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives,* 13(33). Retrieved March 29, 2012, from http://epaa.asu.edu/epaa/v13n33/

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamiltion, L.S. (2003). *Evaluating value-added models for teacher accountability.* Accessed on January 9, 2012, from The RAND Corporation http://www.rand.org/pubs/monographs/2004/RAND_MG158.pdf

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

Mendro, R., Jordan, H., Gomez, E., Anderson, M., & Bembry, K. (1998). *An application of multiple linear regression in determining longitudinal teacher effectiveness.* Paper presented at the 1998 Annual Meeting of the AERA, San Diego, CA.

New Teacher Project. (2007). "Hiring, Assignment, and Transfer in Chicago Public Schools." Brooklyn.

Pearson. (2012). *Evaluating teachers and principals: Developing fair, valid, and reliable systems.* Retrieved from Pearson website: http://educatoreffectiveness.pearsonassessments.com/wp/wp-content/uploads/2012/03/educator-effectiveness-roadmap-03-01-12.pdf

Phillips, S. E. (2009). Legal corner: Using student test scores to evaluate teachers. *NCME Newsletter,* 17(4), 3–6.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods (2nd Ed). Newbery Park, CA: Sage Press.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review,* 94, 247-252.

Ross, S. M., Stringfield, S., Sanders, W. L., & Wright, S. P. (2003). Inside systemic elementary school reform: Teacher effects and teacher mobility. *School Effectiveness and School Improvement,* 14(1), 73–110.

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

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville, TN: University of Tennessee Value-Added Research Center.

Tekwe, C. D., et al. (2004). An empirical comparison of statistical models for valueadded assessment of school performance. *Journal of Educational and Behavioral Statistics,* 29(1), 11–36.

Tucker, P. D. (1997). "Lake Wobegon: Where All Teachers Are Competent (Or Have We Come to Terms with the Problem of Incompetent Teachers?)." *Journal of Personnel Evaluation in Education* 11 (1): 103–126.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. Brooklyn, NY: The New Teacher Project. Accessed February 17, 2012, from http://widgeteffect.org/downloads/TheWidgetEffect.pdf

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education,* 11, 57–67.