Evaluating the Use of Tests to Measure Teacher Effectiveness:

Validity as a Theory-of-Action Framework

Lorrie A. Shepard

University of Colorado Boulder

Paper presented at the annual meeting of the National Council on Measurement in Education

Vancouver, British Columbia

April 14, 2012

The use of student test scores as a basis for evaluating teachers is a relatively new and highly controversial policy. Although systems like the Tennessee Value-Added Assessment System (TVAAS) have been in place in some jurisdictions for more than a decade, Race-to-the-Top funding criteria in 2009 propelled many states to establish systems for evaluating teacher effectiveness taking into account "student growth as a significant factor" (U. S. Department of Education, 2009, p. 9). Some states interpreted this to mean that test-score gains should determine as much as 50 percent of teachers' effectiveness ratings. In this paper, I use the shorthand, test-based teacher evaluation, to refer to a variety of systems in which estimates of teachers' contributions to students' test-score gains play a primary or key role. It should be acknowledged, however, that in most cases student growth data are used in combination with classroom observations and other measures such as parent and student ratings.

Proponents of test-based teacher evaluation argue that growth in student achievement is the ultimate criterion for judging teacher effectiveness. They believe that value-added modeling (VAM) of test-score data can do a better job of identifying the best and worst teachers compared to current indicators and that these methods are sufficiently robust in accounting for initial student differences to provide actionable data (Gordon, Kane, and Staiger, 2006). They point to the inadequacy of input measures such as advanced degrees or scores on licensure exams as indicators of teacher quality (Hanushek & Rivkin, 2006) and to the inability of existing evaluation systems to identify and eliminate bad teachers. VAM detractors claim that neither standardized tests nor VAM's statistical machinery have sufficient validity for the high-stakes purpose of individual teacher evaluation (Baker, Braun, Chudowsky, & Koenig, 2010). These disagreements are more than academic or technical quibbles.

As a result of Race-to-the-Top incentives, high-stakes teacher evaluation systems are being installed in numerous states without adequate validity studies beforehand. Therefore, it is critically important that systematic and rigorous evaluations be conducted of these systems once they are in place. In this paper, I argue that policy intentions should be placed center stage in framing *evaluation studies* – which are essentially *validity investigations*. In making this argument, I first summarize the evolution of contemporary validity theory highlighting in particular its similarity to theory-of-action frameworks, familiar to organizational theorists and policy researchers. I then outline the theories of action underlying the use of tests in support of high-stakes teacher evaluation mandates. In addition to intended effects and claimed benefits, however, validity evaluations also require examination of unintended effects. Therefore, it is necessary to enumerate known criticisms of test-based teacher evaluation and implied negative consequences that should also be investigated as possible outcomes.

The evolution of validity theory, validity argument, and theory-of-action framing

Over the past 100 years, validity theory has become increasingly more encompassing and demanding, requiring that various sources of evidence be brought to bear to defend a particular test use. In the earliest days of testing, a test might be claimed to be valid merely on the basis of a content validity analysis by experts or because of a single correlation with a predicted criterion. Guilford (1946), for example, once famously said that "a test is valid for anything with which it correlates" (p. 429). Today, however, validity evaluations must include both logical and empirical evidence. Moreover the gathering and analysis of this evidence should be organized around the theory of the test, attending to both what it is that the test claims to measure *and* what it is that the testing practice claims to do. The re-framing of validity studies to focus on intended test use happened gradually but is as old as Cureton's (1951) "Validity" chapter in the first

Educational Measurement handbook and Cronbach's (1971) attention to the "decision-oriented interpretation" of a test. Two decades ago, in characterizing this shift over time, I suggested that old validity studies, which asked only "whether the test measures what it purports to measure," were analogous to truth-in-labeling standards, whereas contemporary validity requirements that I and others have argued for are more like the Federal Drug Administration's standards for research on a new drug, requiring that it be shown to be "safe and effective" before it can be released for wide-scale use (Shepard, 1993, p. 426).

Messick (1989) is sometimes cited as if he "added" consideration of the social consequences of tests to the concept of validity, when in fact he merely elaborated and called our attention to a long-standing, fundamental aspect of validity studies that took account of test use. What might be considered new in the 1980s, however, -- in response to concerns about test bias, misuse of intelligence tests for special education placements and court decisions regarding employment tests – was a re-centering of validity studies on intended effects. What did "test use" mean if not a set of claims about how using a test was expected to lead to particular desired outcomes? And once attention was focused explicitly on intended effects of a testing program, it followed inevitably that unintended effects should also be considered – one of many valuable lessons learned from re-conceiving validity research as program evaluation (Cronbach, 1988). Thus an I.Q. test might be sufficiently "valid" for use in a research study to evaluate the long-term effects of fetal alcohol syndrome on cognitive functioning but not be valid for placing children in special education – largely because intended positive outcomes of such placements were not sufficiently great to outweigh negative side effects.

A powerful case in point was provided by the National Research Council Panel on Selection and Placement of Students for Programs for the Mentally Retarded led by Heller,

Holtzman and Messick (1982). The Heller et al. Panel had been convened to examine the causes and possible biases leading to overrepresentation of minority children and males in classes for mentally retarded students. The Panel redefined its charge, however, in a way that exemplifies the re-centering of contemporary validity theory on *the adequacy of a test for achieving its intended outcomes*. In the case of special education placements, the intended outcome was to provide more effective educational interventions – tailored to the student's needs – than would be available in the regular classroom. By asking the larger question as to why disproportion was a problem, the Panel brought into their analysis the bodies of research showing the negative effects of labeling and the poor quality of instruction in segregated special education classrooms. To be valid, they said, an assessment should address a child's functional needs that could be linked to effective interventions. "Thus, assessments can be judged in terms of their utility in moving the child toward appropriate educational goals (p. 99)." The Panel reprised the well-known science emphasizing that I.Q. tests measure current cognitive functioning rather than an inborn trait, but they also made this conclusion beside the point if current functioning could not be matched to effective treatments. Measures of reading comprehension or adaptive behavior had the potential to be more useful, but significantly these measures too could not be claimed to be valid for placement, if placements were shown to be ineffective.

Following Cronbach (1988), Kane (1992), and others, the 1999 *Test Standards* adopted *validity argument* as the framework for organizing and integrating validity evidence. "Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use (p. 9)." Beyond test use and testing consequences, per se, an argument-based approach to validity furthered the refinement of validity theory in two important ways: (1) it required that the theory of the test be

made explicit, with underlying assumptions revealed, and (2) it helped to establish priorities as to which evaluative questions were the most central to investigate. For example, in Kane's (1992) example of an algebra placement test used to assign college students either to a calculus or remedial algebra course, the argument for the test depended on several assumptions, including the assumption that algebraic skills are genuinely prerequisite and used in the calculus course, that the placement test represented those skills well, and that the remedial course was effective in teaching the target algebraic skills. What might not be so obvious was a final assumption that high-scoring students placed directly in calculus would *not* also benefit, i.e. their performance in calculus would not improve, if they too had received the remedial algebra treatment. In other words, differential placement also had to be shown to be effective. Note that Kane (1992, 2001, 2006) has consistently identified two layers to the argument approach, the interpretive argument (which lays out the logic model and substantive claims) and the validity argument (which entails the gathering and analysis of evidence to support the plausibility of the interpretive argument). As a short-hand, the *Test Standards* and others have combined both layers in what is called for in a validity argument approach, and I use that simplification here as well.

Policy researchers likely will recognize a close similarity between validity arguments and theory-of-action frameworks offered by Argyris and Schon (1978) as a means to make explicit individuals' reasoning about how intended causal mechanisms will lead to desired outcomes. Argyris and Schon were particularly interested in studying organizational learning, but in their view, theories of action guide all deliberate human behavior. These theories are essentially socially developed norms or belief systems held by individuals about how the world works. Argyris and Schon further distinguished between *espoused* theories of action and *theories-in-use* inferred from how individuals (or organizations) actually behave. In the case of high-stakes

teacher evaluation, legislative mandates may reflect espoused theories -- where it is asserted that by taking certain actions desired improvements in the quality of teaching will result. In many cases, however, the values, strategies, and assumptions implied by such policies are not called out explicitly. Thus, while Argyris and Schon focused on contradictions between espoused theories and theories-in-use that might hinder organizational learning, a complete analysis of high-stakes teacher evaluation policies requires that we identify both explicit and implicit assumptions about the uses of testing as part of one underlying theory of action.

Although using test results as part of teacher evaluation is relatively new, there is a longer history with test-based accountability policies that can be used to illustrate the analytical power of a theory-of-action framing. For example, Fuhrman (2004) identified the following assumptions comprising the theory of action underlying No Child Left Behind-era accountability systems.

1. *Performance, or student achievement, is the key value or goal of schooling, and constructing accountability around performance focuses attention on it.*
2. *Performance is accurately and authentically measured by the assessment instruments in use.*
3. *Consequences, or stakes, motivate school personnel and students.*
4. *Improved instruction and higher levels of performance will result.*
5. *Unfortunate unintended consequences are minimal.* (pp. 8-9)

Similarly, Baker and Linn (2004) elaborated on an incentives-based accountability model, specifying several enabling conditions that, when compared with the Fuhrman model, make clearer what would be needed to get from step 3 to step 4: a) alternative actions to improve the situation are known and available, b) cognizant individuals and team members possess the

requisite knowledge to apply alternative methods, c) the selected action is adequately implemented, and d) the actions selected will improve subsequent results. These examples are particularly apt and illustrate important categories of assumptions that can be drawn upon in developing a similar chain of reasoning for analysis of test-based teacher evaluation.

Summative and formative theories of action for using tests to evaluate teachers

Teachers make more of a difference in determining the quality of education than any other school-controlled resource. This has always been believed intuitively, but in recent years, the results from studies employing value-added models appear to have demonstrated this empirically. Just as was true 25 years ago at the start of the standards movement, the need for high-stakes teacher evaluation is couched in terms of weak international comparisons and worries about global competitiveness. Today, however, for the first time, extensive test data, data systems (including teacher IDs), and statistical methods are available to permit analyses focused on teachers. Various researchers have quantified the benefit of having a "good teacher" versus a bad one. For example, using three years of data from Los Angeles Unified School District (LAUSD), Gordon, et al. (2006) showed that teachers in the top quartile of value-added effects, during the first two years, produced strong enough gains in year three to give their students a 5-percentile-point benefit over the average teacher and a 10-percentile point over teachers in the bottom quartile of prior effects. It is these documented differences in student growth that test-based, value-added evaluation systems are intended to identify and use in making both formative and summative decisions.

In broad sweep, the general theory of action for test-based teacher evaluation systems holds that using student growth to measure teacher effectiveness will improve the quality of education provided to students and hence will improve student achievement. In this section, I

present the key assumptions for summative and formative versions of this theory, relying where relevant on the same categories and chain of reasoning followed by Fuhrman (2004) and Baker and Linn (2004) for test-based school accountability. These simplified models serve a heuristic purpose. In practice, of course, most teacher evaluation systems will attempt to serve both formative and summative purposes.

Using data similar to Gordon et al. (2006), Hanushek (2011) proposed a quite straightforward summative use for test-based evaluation data: fire the bottom 7-12 percent of teachers and replace them with average teachers. Hanushek calculated that over the 13-year span of one kindergarten-to-12th-grade cycle, the resulting improvement in student achievement would raise the level of achievement in the U.S. to that of Finland. The argument or theory of action underlying Hanushek's proposal and other summative uses can be laid out as follows.

A. *Student achievement is the key value or goal of schooling, and constructing teacher evaluation systems around student growth will focus attention on this valued outcome.* In the past, policies intended to improve teaching quality have focused on inputs, i.e., teacher credentials such as content majors, content tests, and state licensure, but such credentials have not been shown to have a strong relationship with student growth. Why not focus directly on student growth, thus making it possible to "deselect" teachers with the poorest results on student growth measures (Hanushek, 2011)?

B. *Student achievement is accurately and authentically measured by the assessment instruments in use.* The adequacy of assessments to represent intended learning goals is seldom discussed in the VAM literature, although it is sometimes said that measures need only correlate well with achievement targets for value-added models

to work. Calling out this assumption, even if left unstated by various advocates, is consistent with both validity argument and Argyris and Schon's (1978) inferred theories-in-use. Embedded within this general claim regarding achievement tests are more specific beliefs about achievement, assuming for example that math and reading are so fundamental to other learning that these measures stand as an adequate proxy for other important learning goals.

C. *Teacher contributions to growth are accurately quantified by Value-Added Modeling.* The compelling advantage of VAM over previous analytic tools is its promise to control for students' prior achievement and other, out-of-school factors so as to isolate the effects of teaching quality in a given school year.

D. *The poorest teachers can be eliminated on the basis of VAM results and sufficient numbers of teachers with average student growth are available to replace those who are fired.* Hanushek (2011) is aware that implementation of a fire-and-replace strategy would require that districts develop quite different recruitment, pay, and retention policies. Novice teachers are known to have relatively poorer student growth on average, until roughly their third year of teaching (Clotfelter, Ladd, & Vigdor, 2007). To ensure that replacement teachers could, as a group, achieve growth rates commensurate with the district average would require either quite different entry criteria (which Hanushek does not believe are available) or systematic and substantial retention of the better teachers among those who are currently leaving the profession every year.

E. *Improved instruction and higher levels of achievement will result.* In the formative evaluation theory-of-action discussed next there are additional leverage points built

in, showing how instructional practices of teachers across the whole distribution are expected to improve. With the firing-of-bad-teachers approach, teaching practices of individual teachers do not change. Rather, the overall average improves because bad teachers are replaced with better ones.

F. Unfortunate unintended consequences are minimal. Following closely the possible negative side effects identified by Fuhrman (2004), it is assumed that if the evaluation systems work as intended, higher levels of achievement "will not be undermined by perverse incentives or other negative developments" (p. 9). For example, instruction will genuinely improve and not become focused narrowly on test preparation, teachers will attend to all the students in their class and not ignore mobile students who are missing from the VAM database, effective teachers will not decline teaching assignments in under-resourced or low-performing schools, and teachers in the top 90 percent who are not targeted for removal will not feel threatened by test-based teacher evaluation or in other ways feel the quality of their working conditions degraded.

A formative theory of action for test-based teacher evaluation relies on all of the same assumptions as the summative model, except for the replacement assumption. In its place, a formative theory makes assumptions about how the findings and the incentives from the evaluation system will cause existing teachers to improve their instructional practices and thereby increase student growth. The Race-to-the-Top scoring rubrics (U. S. Department of Education, 2009), for example, which induced so many states to adopt test-based effectiveness measures as part of their evaluation systems, gave considerable weight (more than any other single criterion) to the use of student growth to improve teacher and principal effectiveness. Details of the rubric further specified that these evaluations should be used for both formative

and summative purposes. Formative uses included coaching, induction support, and professional development. Summative applications include merit pay, award (or denial) of tenure, and removal of ineffective teachers as described previously. In addition to assessment, data-system, and teacher evaluation specifications, Race to the Top also awarded points for state plans that focused explicitly on providing effective support to teachers and principals that might include common planning and collaboration time for teachers.

> Such support might focus on, for example, gathering, analyzing, and using data; designing instructional strategies for improvement; differentiating instruction; creating school environments supportive of data-informed decisions; designing instruction to meet the specific needs of high-need students; and aligning systems and removing barriers to effective implementation of practices designed to improve student learning outcomes. (p. 10)

To construct the formative theory of action then, we need to add espoused change mechanisms to the previous summative model. In parallel to Fuhrman (2004) and Baker and Linn (2004), these include motivation or incentives for change, means for identifying new and effective practices, and the wherewithal to implement new practices. (I have designated these as D1, D2, and D3 to locate them within the summative chain of reasoning, where they might be used, for formative purposes only, in place of the replacement strategy, or more likely in addition to firing low-performing teachers who fail to improve.)

D1.Consequences, or stakes, motivate school personnel. Accountability pressures generally are intended to increase and focus effort on improvement. Even when the only consequences were publicly-reported, school-level test scores, research on high-stakes testing documented intensive efforts by teachers to raise student performance.

Tying rewards and sanctions for individual teachers to test scores is expected to raise the stakes further.

D2. Individuals and team members have the means to identify new and effective instructional practices. An unfortunate carry over from test-based accountability is the presumption that educators have the means to improve student achievement if they would just try harder. In a recent randomized experimental study of pay for performance, however, even with substantial bonuses there were no achievement gains for treatment classrooms compared to controls, and teachers reported that the chance to win bonuses had not altered their practices "because I was already working as effectively as I could" (Springer, Ballou, Hamilton, Le, Lockwood, McCaffrey, Pepper & Stecher, 2011). Similarly, earlier studies of accountability reforms suggested that only the better positioned school staffs, with sufficient prior knowledge and skills to pursue new curriculum content and new instructional strategies, were able to respond coherently and change their instructional practices productively in response to external accountability pressures (Elmore, 2003). Fortunately, most formatively-oriented test-based teacher evaluation plans have at least acknowledge that new skills are needed, and some have explicit theories about how these might be sought and developed. Race-to-the-Top, for example, adopted a data-based-decision-making theory but assumes that once needs are identified from data, effective interventions would be devised by school staff. The Gates-funded Measures of Effective Teaching (MET) project looks to the other components of the evaluation system beyond test scores, such as classroom observations and student

feedback, as the best means to provide teachers with information about what they need to do to improve (MET project, 2013a, 2013b).

D3. Individuals and team members have the knowledge and support necessary to implement the selected instructional practices. Knowing that a need exists and that a potential solution has been offered is not enough, if the solution requires significant new learning. From the research literature, we know that it is possible to induce greater student learning when challenging curricula, greater student engagement, and higher levels of classroom discourse are instituted, but we also know that it is devilishly difficult to implement such reforms at scale. The Race-to-the-Top theory calls for coaching and professional development, but historically these needed capacity-building supports have not been forthcoming (Elmore, 2003).

Together the summative and formative theories of action constitute the validity argument that can be used to organize evaluation studies of newly adopted test-based teacher evaluation systems. Laying out the underlying logic model or chain of reasoning makes it possible to identify and investigate intermediate steps by which a test-based policy intervention is intended to achieve desired ends. Known shortcomings can also be investigated within this organizing framework.

What do we know so far about Value-Added Models and test-based teacher evaluation systems? What important questions remain?

The point of a validity argument approach is not to elaborate every possible theoretical and empirical thread but rather to focus on the particular claims that are most central to a test use producing its intended outcomes. In this section, key findings from existing research are summarized for each of the theory-of-action assumptions. Past studies do not, however, answer

validity questions for new jurisdictions. Rather they tell us what to watch for in new contexts – especially in the case of unintended effects – and they suggest the kinds of methodologies likely to be fruitful in evaluating test-based evaluation systems. Given that this paper was developed specifically to address the use of tests to evaluate teachers, the analysis of evidence relevant to each assumption is focused primarily on state tests and the use of VAM to attribute achievement gains to individual teachers. Where appropriate, I also comment on the use of additional achievement measures and on the use of classroom observations and student opinion surveys intended to be part of most test-based teacher evaluation systems.

A. Student achievement is the key value or goal of schooling, and constructing teacher evaluation systems around student growth will focus attention on this valued outcome. B. Student achievement is accurately and authentically measured by the assessment instruments in use. These two assumptions are best considered together because despite important distinctions they invoke the same, well-worn research literature on accountability and teaching-the-test effects. The chain of reasoning assumes that accountability (evaluation) pressures will intensify effort, that the right goals have been identified, and that intended goals are represented well enough that test imitation and practice will not somehow cheapen or distort those goals.

There is no question that pressure to improve performance on high-stakes tests works to direct effort -- for good or ill. In countless survey studies from the 1980s to the present, teachers have reported how they shape their teaching practices to conform to the expectations set by end-of-year accountability tests. Negative effects of such pressure are considered under assumption F, such as the elimination of science and social studies from the curriculum along with art, music, field trips, and the like. Positive examples over the last several decades include changes in writing and mathematics instruction in response to test content demands. When state testing

programs added writing assessments, more writing was taught. When mathematics frameworks and assessments were expanded to include geometry and statistics and probability, instruction followed suit. And, importantly for assumption B, when open-ended formats were included asking students to explain their reasoning, students were given more practice explaining, which was followed by improvements in performance (Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996).

Hidden in assumption A is the implied devaluing of aims such as democratic participation, character development, creative expression, ability to work in groups, and the like. Oddly, this is occurring at a time when personal engagement (National Research Council, 2004) and constructs like resilience and grit are increasing recognized to be important for learning. Policymakers and advocates for test-based accountability and teacher evaluation systems may be relying on several different theories-in-use to support this logic. Some believe that reading and mathematics are so fundamental to other learning that these two subjects ought to be given the highest priority even at the expense of other goals; others assume that teachers good at teaching reading and mathematics are also good at creating learning environments that support student development in these other arenas. Unfortunately, what is known from research on high-stakes testing is that attention to other valued goals is disproportionately pushed out of the curriculum in low-performing schools serving poor children. And, from research on learning, we know that decontextualized, test-like, drill-and-practice regimes can actually harm learning, because by removing context they take away the purpose for problem solving and save for later the experience of applying one's knowledge to real-world problems.

Research on the effects of test-based accountability also casts doubt on the claim in Assumption B that student achievement is accurately and authentically measured by current

assessment instruments. It is generally now acknowledged that intensive teaching to low-level tests, fostered by high-stakes accountability policies, leads to curriculum distortion and test-score inflation (Herman, 2004, Shepard, 2008, U. S. Congress Office of Technology Assessment, 1992). Here the concern is not just the neglect of other subject areas but rather a limitation in how even reading and mathematics are taught. All of this is expected to be remedied, of course, with the adoption of more challenging Common Core standards and the development of "next-generation" assessments by the PARCC and Smarter Balanced assessment consortia. The central evaluation question will be to test whether these promised changes are realized. When teachers redirect instructional effort to focus on improving performance on the new tests, are they supporting deep learning? And do apparent learning gains generalize to independent measures of the same knowledge and skills? In past studies of accountability testing, for example, impressive gains on state tests were not always corroborated by state results on the National Assessment of Educational Progress (Klein, Hamilton, McCaffrey, & Stecher, 2000). And because of past evidence showing that poor children in poor schools are disproportionately the victims of test-driven curricula, evaluation studies should be designed with this equity issue in mind. These measurement-focused questions should also be linked to studies examining Assumptions D2 and D3, as we know from past research that superficial efforts to raise test scores often occur in schools that lack a knowledge base and support to help teachers learn to teach in fundamentally different ways.

Teacher surveys will continue to be a worthwhile component in studies examining the validity of the evaluation system. Although of course self-interested, carefully gathered and anonymous survey data can provide important insights about the kinds of instructional decisions teachers are making in response to testing mandates. And in keeping with lessons from test-

based school accountability, at least some investment should be made in an audit test to check on the validity and credibility of gains on tests to which incentives are attached (Koretz, 2003). The recently reported Gates Measures of Effective Teaching (MET) study is exemplary in this regard for having included other measures of reading and mathematics achievement beyond the state tests, which were the primary means for determining value-added estimates of teacher effectiveness. MET researchers selected the Balanced Assessment in Mathematics (BAM) and Stanford 9 (SAT9). Both measures use open-ended item formats and require students to explain their reasoning, which means they are more conceptual and more likely to tap higher-order thinking skills than traditional multiple-choice tests. When these more conceptual tests were used in place of state tests to obtain VAM estimates of teacher effectiveness for the same classrooms of students the correlations were small to modest, .22 in the case of the two reading measures and .38 for mathematics (MET project, 2010). One possible explanation for these relatively weak correlations is that instruction aimed at making gains on the state test is not the same as instruction leading to deeper understanding, hence the finding for large numbers of teachers that gains on state tests did not generalize to the more conceptual tests.

C. Teacher contributions to growth are accurately quantified by Value-Added Modeling. The whole point of VAM is to "level the playing field" so as to make fair comparisons among teachers. Policymakers and educators understand that that raw achievement test scores tend to rank schools by the socio-economic status of the students served. The very name, "value-added," reflects the desire to isolate the unique contribution of schools or teachers to achievement outcomes. The question is, are the statistical adjustments sufficient to accomplish this purpose? Or, stated another way, are the statistical adjustments able to isolate the causal

effect of individual teachers on student achievement and to separate these teacher effects from other contributing factors?

Using both panel data from North Carolina and simulations, Rothstein (2010) developed falsification tests for several prominent value-added models. Essentially he applied the analyses in reverse to see if including indicator variables for fifth-grade teachers would produce the hypothesized zero relationship with fourth-grade student achievement. Instead he found substantial effects of fifth grade teachers on fourth-grade test score gains, which he attributed to non-random assignment of students to classrooms (tracking) and to what he called reversions. Going in the fourth-grade to fifth-grade direction, VAM's less than adequate control means that fifth-grade teachers receive unfair credit when they are assigned students who underperformed in fourth grade who then *revert* to a higher rate of gain in fifth grade. He demonstrated the magnitude of this effect by showing that the residuals from grade four and grade five gains were serially and inversely correlated, -.38 in math and -.37 in reading. Rothstein (2010) warned from these findings that "policies based on these VAMs will reward or punish teachers who do not deserve it and fail to reward or punish teachers who do" (p. 211). He argued that the stakes attached to VAM measures of effectiveness should be relatively small, and cautioned that incentives from a VAM-based system would induce teachers to seek students with the likelihood of good value-added scores.

In contrast to Rothstein's evidence of invalidity, Chetty, Friedman, & Rockoff (2011) demonstrate long-term effects such as a 1% benefit to students' lifetime earnings of having a teacher whose value-added is one standard deviation above the mean. Chetty et al. also provided compelling evidence showing that teachers in the top 5% on VA who changed schools had clear and predictable effects on their schools' grade-level VA, lowering VA for the schools that they

left and raising VA in their new schools. Inverse patterns were found for the bottom 5% of teachers entering and leaving a school. It would be hard to argue that such findings could obtain is there were not some "real" component to VAM effectiveness measures. Of course, it is quite likely that every teacher's result would have both some real and some systematic but untrue component and that the relative size of each would vary by both educational setting and data structures.

As Rothstein (2010) suggests, "any proposed VAM should be subjected to thorough validation and falsification analyses" (p. 210). The likelihood that student characteristics are being falsely attributed to teacher effects should be systematically examined. For example, beyond the general idea of non-random assignment of students to classrooms, are teachers with higher numbers of special education students or English Language Learners more often classified as ineffective? Amrein-Beardsley and Collins (2012) find evidence of such biases in individual cases. For example, one teacher's ranking changed from the bottom to the top decile in the space of two years, when the proportion of ELL and low-income students in her class was dramatically reduced. Although single cases do not help to resolve validity questions one way or the other, they are suggestive of the kinds of trends that should be checked for systematically. Are there patterns like this across cases whereby student characteristics are associated either with rankings or changes in ranking? Other patterns to watch for are floor and ceiling effects on the test, which could reduce VAM effects for teachers serving self-contain classrooms of very low-performing or very high-performing students?

When systematic patterns are found, it will be important to try to understand them, not necessarily just to correct them with further statistical adjustments. In a recent study with Missouri data, Ehler, Koedel, Parsons, and Podgursky (2013) demonstrated that both Student

Growth Percentiles (Betebenner, 2009) and a simple VAM model still resulted in moderate correlations between school effects and measures of student poverty, -.37 and -.25 respectively. Ehler et al. argued for further corrections so that the same proportion of schools would be found effective within each stratum of schools. This would keep the incentives right, they said, rewarding those who were doing relatively well among high-poverty schools and not falsely crediting schools doing less well than others among affluent schools. They acknowledged, however, that they were not necessarily making the best decision in terms of causal inference, given that labor market options could genuinely cause there to be less good teaching in high-poverty schools. Obviously, more of this kind of thinking needs to be done regarding models that over- and under-correct, both for making decisions about individual teachers and when evaluating validity in relation to other measures of effective teaching.

One of the greatest worries about the validity of VAM results is year-to-year instability in estimated teacher effects. Are VAM estimates reliable enough such that teachers are classified with some minimum level of consistency? One of the most widely quoted studies by McCaffrey, Sass, Lockwood, and Mihaly (2009), for example, found correlations from one year to the next ranging from 0.22 to 0.46 for elementary teachers and 0.28 to 0.67 for middle school teachers. Much, but not all, of this instability in estimates of value-added can be attributed to chance in the way that cohorts of students are assigned to teachers from year to year. Because student-specific error tends to cancel out as more students are added to a class, greater stability was found for middle schools teachers serving larger numbers of students than for their elementary school counterparts. One way to quantify the practical significance of instability in VAM estimates is to consider the consistency of normative teacher effectiveness classifications from year to year. For example, in a study of Chicago teachers, Aaronson, Barrow and Sander (2007) found that only

36 percent of teachers ranked in the bottom quartile of teachers in year one of the study remained in the bottom quartile the second year; 35 percent of the bottom-quartile teachers had move to the top half of the distribution by year two. Imagine trying to use a measuring instrument or policy device when results will be seriously wrong one-third of the time. While individual teachers might be expected to have a good year or a bad year (accounting for some small portion of the year-to-year change), the validity of an effectiveness measure logically requires that it detect some persistent teaching quality construct. The whole point of test-based teacher evaluation is to identify enduring effectiveness characteristics of teachers who can then be appropriately selected or rewarded.

As McCaffrey et al. (2009) have shown one solution to annual instability is to average VAM estimates from at least two years. A remedy such as this is certainly a minimum requirement for any high-stakes use of VAM results. Another safeguard would be to bracket (and treat as problematic) any VAM results based on fewer than 15 students. When VAM results are used for formative purposes, however, year-to-year instability creates a quite different problem. Here the conversation about the interpretation of results is likely to occur based on only one year of data (unless results are literally withheld until two years of data have been analyzed). Wide fluctuations as well as individual results that lack face validity are likely to be visible to teachers within a school and could well undermine the trust and credibility needed for effective formative reflection and improvement. It would be wise, therefore, to acknowledge the possibility of estimation error and to triangulate with other indicators of effectiveness as discussed in a later section.

Year-to-year consistency and the adequacy of prior achievement corrections are critical prerequisites, but the real test of validity – as to the construct being measured -- requires

corroborating evidence from independent measures of effective teaching. The MET study, cited previously, was designed to provide this type of convergent validity data by collecting classroom observations, student ratings, and VAM estimates for a large sample of teachers. Mihaly, McCaffrey, Staiger, and Lockwood (2013) found "evidence that there is a common component of effective teaching shared by all indicators" (p. 2). However, the cross-method correlations were distressingly weak, and the researchers rightly noted that there were also substantial differences across measurement modes. The state value-added measure correlated on the order of .17 to .42 with classroom observations and only slightly better with the more reliable student survey measures, .11 to .57. The convergence of evidence appeared to be best for measures of middle school math teaching, where none of the cross-method correlations fell below .35. Strong method-specific variance was also evidenced by the within-method correlations. Observational indicators correlated with each other on the order of .53 to .99 and student survey indicators were correlated on the order of .69 to .94. High within-method correlations compared to the between-method correlations are troubling if the hypothesis is that these methods measure a single, common construct.

Because the MET researchers are focused on creating the most defensible operational system they go on to develop composites that combine the reliability of student ratings, the improvement information in classroom observations, and the policymaker's interest in student outcomes. However, validity researchers should be prepared to look further in trying to explain the off-diagonals in scatterplots of weak correlations like those found in the VAM study. With state-test VAM on the x-axis and alternative-test VAM on the y-axis, teachers in the upper-right-hand quadrant are unambiguously good on both, but what about the contradictory quadrants? What if we gathered independent evidence of teachers' emphasis on test preparation? Would

high scores on this new indicator account for a disproportionate number of teachers in the lower-right quadrant? And what kind of teachers land in the upper-left-hand quadrant? Are affluent or honors classrooms represented here due to ceiling effects on the state test? Surely we want to answer these types of questions in evaluation studies -- where additional information can be sought -- before settling on decision rules with only the state test data available.

A 2011 study by Hill, Kapitula, and Umland provided the kind of explanatory validity investigation that is needed. Like the MET study design, Hill et al. included other measures of teaching quality in middle school mathematics classrooms and framed their analyses using a validity argument approach. Value-added effects in mathematics were determined using the state test for the entire district sample of teachers, but the other measures were administered to a small subsample of teachers to enable in-depth analyses. These measures included a nationally developed survey of teachers' mathematical knowledge for teaching (MKT) and an observational protocol focused on the mathematical quality of instruction (MQI). A separate, observational measure of MKT was also created to reflect the level of mathematical knowledge for teaching reflected in observed lessons. Figure 1 from Hill et al. (2011) displays the relationship between VAM effects and scores on the MQI for the 24 teachers in the in-depth study. The different teacher symbols denote four fictionally-named schools.

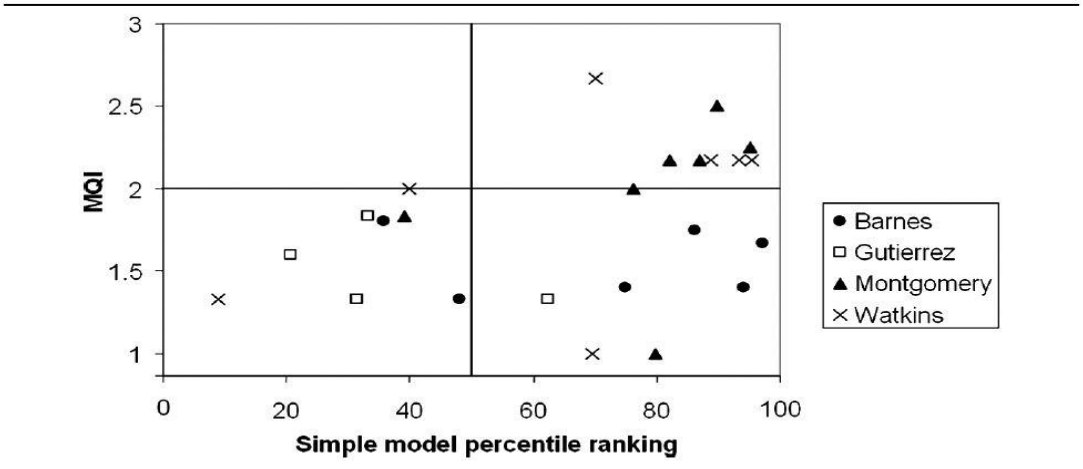

Figure 1. **Mathematical quality of instruction (MQI) versus value-added score from the simple model.**

[I will need to get permission from AERJ and Hill to reproduce Figure 1.]

In the Hill et al. (2011) study there were no "false negatives," that is, teachers who scored very poorly on the value-added measure but who did well on the measure of quality mathematics instruction. This finding would not necessarily generalize to other measures and settings but is exactly the kind of question that must be investigated for each evaluation system. The power of the Hill et al. study is best illustrated, however, by focusing on the "false positives," in particular the two teachers who performed very well on the value-added measure but who scored at the very bottom of the MQI (scores of 1.0). Hill et al. developed detailed case studies describing the nature of mathematics instruction across six observations. In the first case, with VAM scores at the 70th percentile district wide, significant problems occurred with basic mathematics in all of the lessons observed.

> (The teacher) reasons incorrectly about unit rates. She concludes that an answer of 0.28 minutes must actually be 0.28 seconds because one cannot have a fraction of a minute. She tells students that integers include fractions. She reads a problem out of the text as 3/8 + 2/7 but then writes it on the board and solves it as 3.8 + 2.7. She calls the commutative property the community property. She says proportion when she means ratio. She talks about denominators being equivalent when she means the fractions are equivalent. (Hill et al., 2011, p. 27)

In the second case with VAM scores at the 80th percentile district wide, limitations in the quality of teaching were more like the other false positive cases with very little mathematics instruction going on, relying instead on assigning problems from the book. The authors concluded that it was more likely the capabilities of the accelerated and highly motivated students that produced

high VAM results rather than the quality of mathematics teaching that had produced the high VAM scores for this teacher. From a validity argument perspective, Hill et al. concluded that value-added analyses—at least for this particular specification—did not have validity for identifying and rewarding effective teaching. They argued that the public nature of value-added-based rewards was likely to create distorted incentives with teachers competing to teach accelerated students and declining to teach special education students, who in the larger data set appeared to depress teachers' VAM results.

To be fair, alternative measures of teaching quality also have significant drawbacks. Classroom observation methods are costly and require observations across multiple occasions. As part of the MET study, Ho and Kane (2013) documented, for example, that single observations have a reliability of only .27-.45 and having at least two observers was the single best way to improve reliability. Their most telling finding, however, was that observational ratings were undifferentiated and constrained to the middle of the score scale, with raters rarely scoring either a one or a four on a four-point score scale. Given that "distinguished" ratings were even less frequently awarded than "unsatisfactory" ratings and much of the rating was done by external raters, this restricted-range phenomenon cannot be explained by the familiar complaint that administrators unduly score all their teachers as satisfactory. It suggests rather than clearly deficient practices are extremely rare, when aggregated across dimensions of the observation protocol, and should be taken seriously when they do occur. For the purpose of validity research, it will be important to test the adequacy of observational measures in contexts where distinguished practice is known to exist; otherwise, it is impossible to know whether it is the teachers or the instruments that are limited. Student ratings are the most straightforward and reliable indicators of teaching quality, but they have the potential to reward popularity over

learning goals and have been associated with grade inflation in higher education (Eiszler, 2002; Germain & Scandura, 2005).

Given the fallibility of each of the measures of effectiveness, it will be important to triangulate evidence. Composite measures that merely weight and add separate indicators are not the same as triangulation, which involves not only recognizing agreement among measures but also consideration of the likely explanation of inconsistencies. To be sure, at the bottom of the scale, a teacher who is low on all indicators (VAM plus observations and parent surveys) has very little explanation or sympathy, especially if it is the case that unsatisfactory observational ratings are given rarely. However, two teachers with composite scores placing them at the 25th percentile could be two quite different cases. One might have deficient classroom practice and poor student ratings but be "rescued" somewhat by VAM scores similar to the false positive cases in Hill et al. (2011). The other might have excellent classroom practice and enthusiastic student ratings but have poor VAM scores. I would be inclined to consider the second teacher to be the more effective teacher, if I also knew that she had a disproportionate number of special education students and that percent of special education students was known to be inversely correlated with VAM in my district. Clearly, judgment will be required to make sense of both aggregate and individual teacher data. While it may not be cost effective to analyze every individual case in this way, evaluation studies should look for regularities in these types of analyses, and of course individual cases singled out for high-stakes decisions warrant this kind of attention.

D. The poorest teachers can be eliminated on the basis of VAM results and sufficient numbers of teachers with average student growth are available to replace those who are fired. D1.Consequences, or stakes, motivate school personnel. D2. Individuals and team members

have the means to identify new and effective instructional practices. D3. Individuals and team members have the knowledge and support necessary to implement the selected instructional practices. In both the summative and formative theories of action, the D-step is the action step. On the basis of VAM test results and other indicators, low performing teachers are either to be fired or improved.

Validity evidence for VAM alone is insufficient to warrant firing even with two years of data because of the regularities by which teachers are assigned the same types of students across years. However, *with confirmatory evidence* from multiple measures, removal of Hanushek's (2011) very bottom percent of teachers could possibly be defended. Targeting the tail of a composite with confirming evidence from each of several measures means that decisions are more demonstrably valid. Keeping in mind, however, that only about 4 percent of teachers were scored as unsatisfactory by trained observers in the MET study (Ho & Kane, 2013), it would be difficult to justify percentages as high as those proposed by Hanushek. Removing only extreme cases also makes it more likely that even a replacement novice teacher will be able to perform better. The difference in value-added effects between novice teachers and teachers with five or more years of experience has been widely documented (Clotfelter, Ladd, Vigdor, & Diaz, 2004; Hanushek, Kain, & Rivkin, 1998), and is a large effect compared to other VAM estimated effects, on the order of one-tenth of a standard deviation. Although replacing fired teachers with novice teachers will likely not have as great a benefit as Hanushek's hypothetical substitution of average teachers, it is a much more realistic projection, and despite the lower performance of novice teachers is likely to be an improvement on average if only the very worst, say 4 percent of teachers are replaced by novices. Rhetorical restraint regarding the percentage of teachers who could reasonably be identified and fired would also quiet some of the anxiety associated with

implementation of test-based evaluation systems. Obvious examples of invalid attributions will make it very difficult for teachers to trust the system and be willing to learn from the data; hence the need to reassure participants that safeguards are in place to verify the reasonableness of firing decisions.

The formative logic model parallels that of existing school-level accountability models. It assumes that attaching consequences to test results will motivate educators to try harder, that they will be able to identify more effective instructional practices and will have the supports necessary to enact those practices. In my earlier outlining of the model, I summarized previous research evidence suggesting why these assumptions are not likely to hold true, especially in low-performing, under-resourced schools. Here, it might be helpful to mention other bodies of research that should be called upon in the future to frame evaluations of test-based teacher effectiveness systems. These research literatures include research on motivation, teacher professional development, high-leverage instructional practices, and curricular coherence. Economists who are promoting the use of VAM do not have an idea about what substantive changes they are trying to induce in classrooms. Ehler et al. (2013), for example, rightly worry that the better teachers in low-performing schools will be discouraged if they are not allowed a relative standard of comparison, but what if they are achieving relatively good results by energetic drill and test preparation (producing the kinds of gains that don't generalize to the conceptual measures in the MET study)? These suspicions, or call them hypotheses, are what should be investigated as these VAM reward structures are brought on line.

A recent National Research Council (2011) report, *Incentives and Test-Based Accountability in Education,* provides a useful summary of motivation research related to learning as well as research on incentives and job performance in the economics literature.

Several key points are worth mentioning. Rewards (such as performance pay) can increase or decrease motivation depending on whether they are felt to signify competence or are perceived as controlling. Feedback focused on task features (such as increasing wait time) improves performance, whereas normative comparisons (like VAM) can reduce motivation. Unrealistically high goals are demotivating. Public servants are primarily motivated by internal rewards – trust, autonomy, job satisfaction, and goals of the organization. Similarly, there is an extensive body of research on teacher learning and teacher professional development. As summarized by Darling-Hammond (2008), effective professional development is sustained and intensive, supported by modeling, coaching, and problem solving directly tied to dilemmas that arise in enacting change. Effective professional development supports teachers in looking closely at student work and student thinking, involves collaborating with other teachers, and is coherently linked to other aspects of school change. Some of these ideas were invoked in the Race to the Top intentions, but they are obscured in many VAM applications focused only on summative conclusions and actions. The example of Los Angeles teachers being publically ranked by the *LA Times* (Felch, Song, & Smith, 2010) is a case in point.

Surely, in an environment of effective and ineffective individual ratings, all teachers will be scurrying to improve. The danger is that ill-directed effort could lead to many superficial and disconnected changes without any real improvement in the quality of instruction. More than a decade ago, researchers in Chicago examined the problem of "Christmas tree" innovations that resulted from dozens of project, programs, and partnerships being adopted in hopes of reform. By contrast, schools that developed more coherent instructional frameworks saw much greater improvement in student achievement. Instructional coherence was marked by a common framework that guided curriculum, teaching, assessment, and learning climate, professional

development and accountability structures that supported use of the shared framework, and sustained resources and teaching assignments to help teachers learn how to teach well in their specific roles (Newmann, Smith, Allensworth, & Bryk, 2001). While some attention to actual classroom practice is intended by use of observation protocols in educator effectiveness systems, care will be needed to make sure that the process can actually be used in a way that supports teacher learning. Teachers need strategies and opportunities to work on specific high-leverage instructional practices such as leading a discussion -- which includes developing generative questions and revoicing or extending student ideas (Grossman, Hammerness, & McDonald, 2009) – but they also need coaching and supported practice to see how specific strategies are part of a coherent whole.

E. Improved instruction and higher levels of achievement will result. F. Unfortunate unintended consequences are minimal. Assumptions E and F refer to the outcomes of the theory of action, both intended and unintended. Improving instruction is an intermediate outcome expected, in turn, to improve student achievement. Both links in the causal chain should be evaluated. MET researchers Ho and Kane (2013) have provided an insightful treatment of the sources of unreliability in observers' ratings of classroom practice. To account for real differences in VAM results (the replicable portion of findings like those in the Gordon et al. (2006) study, they called for more research on instruments that can better discriminate among teachers and they cited in particular efforts to develop subject-specific measures such as the MQI used in the Hill et al. (2011) study. The video-taped lessons from the MET study are a significant resource and will enable after-the-fact comparisons of the predictive utility of new measures of teaching practice. More importantly this technology should be used again to establish baseline data in districts and states implementing educator effectiveness systems to

ensure that real changes in practice can be evaluated over time and be distinguished from observation-scale inflation.

Test-based teacher evaluation systems are as yet too new to permit meaningful analyses regarding the effects of such systems on student achievement. The Tennessee Value-Added Assessment System (TVAAS) (Sanders & Horn, 1994) is two decades old, but very little evaluative information exists as to how it affected instructional practices or personnel structures in Tennessee. Nor are there carefully controlled analyses addressing specifically whether NAEP gains have been better in Tennessee than elsewhere in the nation. Simple inspection of Math and Reading NAEP data from 1992 to 2011 suggests that Tennessee has tracked below the nation by roughly the same amount over the entire 20-year period. Test-based school accountability has produced an extensive body of sophisticated analytic methods designed to parse the unique effects of accountability policies on NAEP test score gains. This research is summarized in the NRC (2011) Incentives report and in Dee and Jacobs (2011). It presents a mixed picture of small but statistically significant gains in 4th-grade mathematics reasonably attributable to accountability incentives but essentially no gain in other grades and subjects. These same kinds of analyses will be needed over time to assess the effects of test-based teacher evaluation systems on achievement. Widespread concomitant changes in instructional efforts prompted by adoption of Common Core State Standards and new assessments will pose additional attribution challenges. However, if sufficient numbers of states elect not to use VAM to make high-stakes decisions about teachers, a natural experiment will be underway that would enable meaningful comparisons (Rubin, Stuart, & Zannato, 2004).

Assumption F, of course, makes the unlikely claim that there will be no negative side effects, though surely there will be some. The question is how serious are unintended effects?

We already have some warnings that will need to be investigated systematically. The 2012 *MetLife Survey of the American Teacher* (Markow & Pieters, 2012) reported the lowest level of teacher job satisfaction in 20 years, dropping 15 percentage points since 2009, from 59% who were very satisfied to 44%. Those likely to leave the profession has increased from 17% to 29%, those not secure in their jobs has increased from 8% to 34%. Teachers with lower job satisfaction are also less likely to say that they are treated as professionals by the community, 68% versus 89%. While these trends are more likely the cumulating effects of NCLB, such findings do not bode well for further test-based accountability pressures and should surely be studied in comparisons between high- and low-stakes teacher evaluation jurisdictions.

In general the search for plausible unintended effects falls into a few broad categories. First, as in the example above, what happens to the teaching workforce or teaching profession? Are teachers especially those with high VAM scores more or less likely to stay in the teaching profession? Are effective teachers better distributed to high-needs schools than occurred prior to system implementation? Over the next decade, are college students more or less likely to view teaching as a desirable profession? A second category of effects has to do with the many variants on the teaching-the-test theme. Are new assessments significantly improved so that test practice leads to real gains that can be confirmed by a low-stakes audit test (thereby verifying the credibility of dramatic gains at least for the system as a whole)? Just as extra attention to "bubble-kids" was associated with NCLB's incentives tied to moving students over the proficiency cutscore threshold, does VAM create new incentives to ignore students whose likely gains are out of range of the test or who are new to the school and therefore lack pretest data?

A final category for evaluating plausible unintended effects has to do with the availability of resources to help teachers improve. Race to the Top's theory of action included expectations

that coaching, induction support, and professional development would be part of new high-stakes systems. In an effort to examine the effects of NCLB on necessary intermediate variables, Dee and Jacob (2011) used an interrupted time series design and found that, indeed, NCLB leveraged meaningful increases in resources, such as increased teacher compensation and the proportion of teachers with graduate degrees. Research by Elmore (2003) and others warns, however, that promises for capacity building are not always kept. Similarly in the context of teacher evaluation, the allocation of resources should be studied as an outcome of high-stakes systems in addition to being an important mediating variable for understanding effects on teaching practice and student achievement.

Conclusion

New, high-stakes systems for evaluating teachers -- using test-based, value-added estimates of student growth as a primary ingredient -- have been developed and implemented without conducting validity evaluations beforehand. Given the stakes involved and the potential for systematic biases in estimates of teachers' contributions to student growth, it is essential that test-based teacher evaluation systems be rigorously evaluated. This paper provides a brief summary of contemporary validity theory -- which focuses on the adequacy of a test for achieving its intended purposes. A validity argument, used to frame and organize a validity evaluation, is analogous to theory-of-action frameworks familiar to policy researchers. In essence a validity investigation requires that evidence be gathered to examine whether intermediate steps in the logic model are functioning as intended and to document consequences, both intended and unintended.

For test-based teacher evaluation, the underlying chain of reasoning or logic model assumes that achievement tests are sufficiently robust to represent student learning well for the full range of achievement and that VAM statistical adjustments are discerning enough to

disentangle the effects of individual teachers versus home resources, past teachers, and current classmates. The intended summative purpose of such systems is to fire bad teachers and replace them with better teachers, with the expectation that the quality of instruction will improve producing, in turn, greater gains in student achievement. The formative argument assumes that test-based evaluations will create both the information and the incentives needed to help teachers improve. It is also assumed that teachers with poor effectiveness ratings will have the support they need to identify and implement new and effective instructional practices. Validity evaluations require that these intended mechanisms and effects be investigated and, at the same time, that plausible unintended effects be examined.

The existing research literature on VAM cannot give a one-time answer that VAM is or is not valid for teacher evaluation. Rather, the validity of teacher effectiveness ratings in any given state or district will depend on several factors: on the particular achievement measures used to assess the outcomes of learning, on the adequacy of prior achievement data, on the assignment of students to classrooms, on the concurrent effects of other learning resources, on the particular VAM specifications, on the quality of observational and other measures of effectiveness used in the system, and on the judgments involved in weighing evidence from multiple measures. At best, existing research offers insights about the potential threats to validity that need to be addressed.

Each of the indicators in a multiple measures system – classroom observations, VAM scores, and student ratings – have sources of error due to both unreliability and bias. Thus it will be essential to triangulate and look for congruence of evidence. When indicators diverge, explanations should be sought rather than merely averaging the results. Correlational scatterplots can be used strategically to frame subsequent validity steps. Cases in the "off-

diagonals," teachers for example who are very high on classroom practices and very low on VAM, should be examined for patterns in that particular jurisdiction. When systematic disproportions occur, such as over-representation of special education or honors teachers in a contradictory quadrant, then these patterns should be considered with making high-stakes decisions about individual cases. Past research offers possible explanations for contradictory results that may reflect systematic biases that bear on the validity of effectiveness ratings -- teaching the test, disproportionate numbers of ESL students, or floor and ceiling effects on the test.

Overtime the effects of the teacher evaluation system must be evaluated in relation to its intended impacts on teaching and learning. What changes in teaching practices are reported by teachers and documented by observational measures and student ratings? What changes occur on high-stakes achievement tests compared to the baseline year, and are these effects confirmed by independent audit tests? Knowing what is already known from past research on accountability structures, major categories of likely side effects to be investigated include gaming, effects on the teaching workforce, and differential availability of resources to improve. Ultimately a test use and the system in which it is embedded can only be said to be valid if they lead to genuine improvement in the educational system.

References

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

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

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS EVASS) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives*, *20*(12), 1-27.

Argyris, C., & Schon, D. (1978). *Organizational learning: A theory of action perspective*. Reading, MA: Addison Wesley.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*, EPI Briefing Paper #278. Washington, DC: Economic Policy Institute.

Baker, E. L., & Linn, R. L. (2004). Validity issues for accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 47-72). New York: Teachers College Press.

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

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: The National Academies Press.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*, Working Paper 17699. Cambridge, MA: National Bureau of Economic Research. http://www.nber.org/papers/w17699

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects, *Economics of Education Review, 26*, 673-682.

Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do school accountability programs make it more difficult for low-performing schools to attract and retain high quality teachers? *Journal of Policy Analysis and Management*, *23*(2), 251–271.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun, Eds., *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Curton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington, DC: American Council on Education.

Darling-Hammond, L. (2008). Teacher learning that supports student learning. In B. Z. Presseisen (Ed.), *Teaching for intelligence,* Second Edition (pp. 91-100). Thousand Oaks, CA: Corwin Press.

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

Ehlert, M., Koedel, C., Parsons, E., Podgursky, M. (2013). *Selecting growth measures for school and teacher evaluations: Should proportionality matter*? CALDER Working Paper No. 80. Washington, D.C. CALDER, American Institutes for Research.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, *43*(4), 483-501.

Elmore, R. (2003). Accountability and capacity. In M. Carnoy, R. Elmore, & L. Siskin. (Eds.). *The new accountability: High schools and high-stakes testing*. New York: Routledge Falmer.

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

Fuhrman, S. H. (2004). Introduction. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 3-14). New York: Teachers College Press.

Germain, M., & Scandura, T. A. (2005). Grade inflation and student individual differences as systematic bias in faculty evaluations. *Journal of Instructional Psychology, 32*(1), 58-67.

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

Grossman, P., Hammerness, K., & McDonald, M. (2009). Redefining teaching, re-imagining teacher education. *Teachers and Teaching: Theory and Practice, 15*(2), 273-289.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*,427-439.

Hanushek, E. A. (2011). Valuing teachers. *Education Next, 11*(3), 41-45.

Hanushek, E.A., Kain, J.F., Rivkin, S.G. (1998). *Teachers, schools, and academic achievement, NBER working paper series*, Working Paper 6691. Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education*, Vol. 2. Amsterdam, The Netherlands: Elsevier B. V.

Heller, K. A., Holtzman, W. H., & Messick, S., Eds. (1982). *Placing children in special education: A strategy for equity*. Washington, DC: National Academy Press.

Herman, J. L. (2004). The effects of testing on instruction. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 3-14). New York: Teachers College Press.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794-831.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Bill & Melinda Gates Foundation.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

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

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill & Melinda Gates Foundation.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., Stecher, B. M. (2000). *What do test scores in Texas tell us*? Santa Monica, CA: RAND.

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice*, *22*(2), 18-26.

Markow, D., & Pieters, A. (2012). *The Metlife survey of the American teacher: teachers, parents and the economy*. New York: Metlife.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

MET project. (2013a). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Seattle, WA: Bill & Melinda Gates Foundation.

MET project. (2013b). *Feedback for better teaching: Nine principles for using measures of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation.

MET project. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching project*. Seattle, WA: Bill & Melinda Gates Foundation.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation.

National Research Council. (2011). *Incentives and test-based accountability in education*. M Hout & S. W. Elliott (Eds.), Board on Testing and Assessment. Washington, DC: The National Academies Press.

National Research Council and the Institute of Medicine. (2004). *Engaging schools: Fostering high school students' motivation to learn*. Washington, DC: The National Academies Press.

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). Instructional program coherence: What it is and why it should guide school improvement. *Educational Evaluation and Policy Analysis, 23*(4), 297-321.

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

Rubin, D., Stuart, A., & Zannato, E. (2004). A potential outcome view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103-116.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*(3), 299-311.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education*, Vol. 19 (pp. 405-450). Washington, DC: American Educational Research Association.

Shepard, L. A. (2008). A brief history of accountability testing, 1965-2007. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability*. New York: Routledge.

Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (1996). Effects of Introducing Classroom Peflormance Assessments on Student Learning. *Educational Measurement: Issues and Practice*, *15*(3), 7-18.

Springer, M. G., Ballou, D., Hamilton, L., Le, V., Lockwood, J. R., McCaffrey, D. F. Pepper, M., & Stecher, B. M. (2011). *Teacher pay for performance: Experimental evidence from the project on incentives in teaching*. Society for Research on Educational Effectiveness.

U. S. Congress Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U. S. Government Printing Office.

U. S. Department of Education. (2009, November). *Race to the Top program executive summary*. Washington, DC: Author.